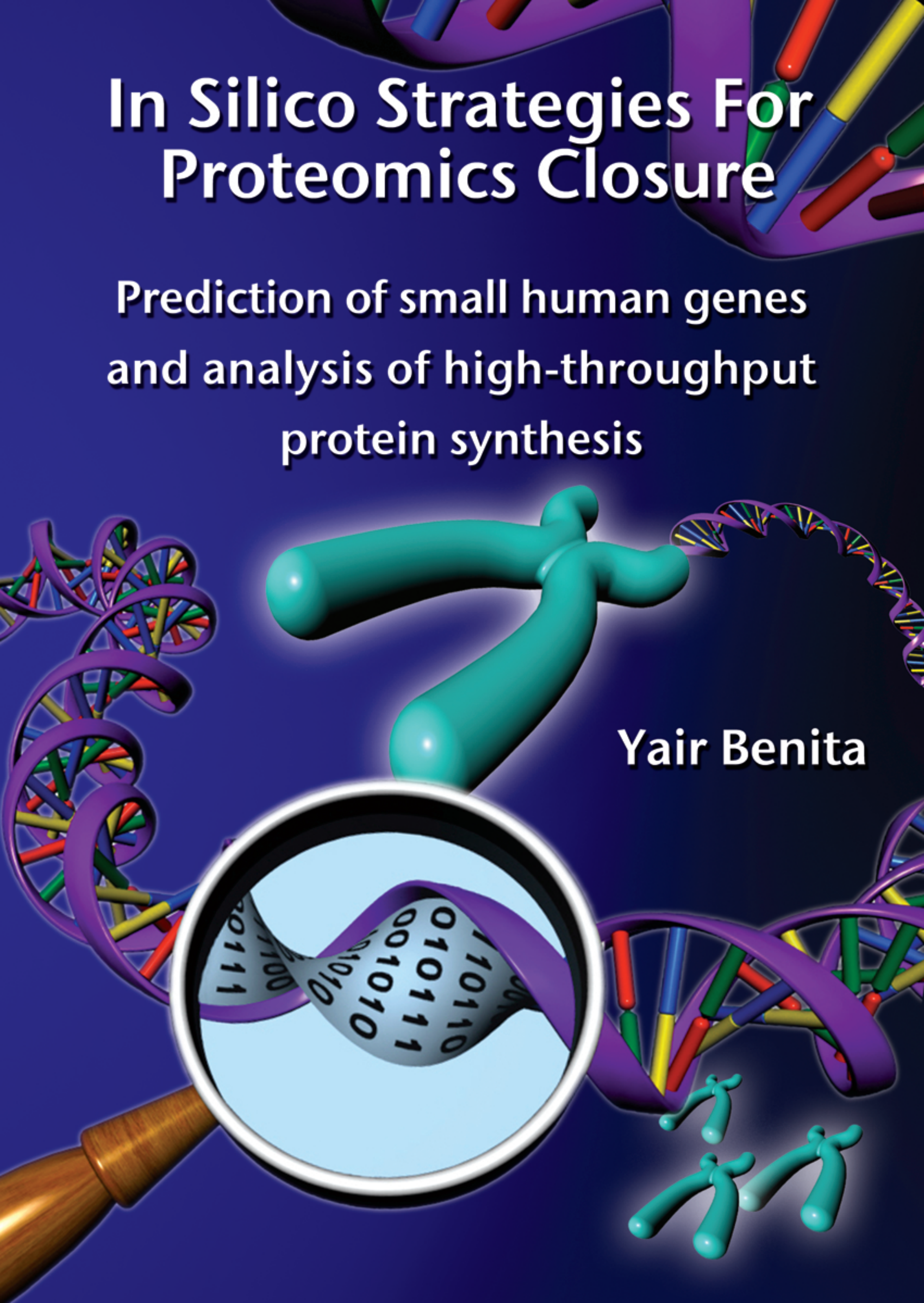
In Silico Strategies For Proteomics Closure

Prediction of small human genes and analysis of high-throughput protein synthesis

Yair Benita



In Silico Strategies For Proteomics Closure

Prediction of small human genes and
analysis of high-throughput protein synthesis

Yair Benita

Yair Benita

IN SILICO STRATEGIES FOR PROTEOMICS CLOSURE: Prediction of small human genes and analysis of high-throughput protein synthesis

Thesis, Utrecht University, 2006

ISBN-10: 90-9020497-0
ISBN-13: 978-90-9020497-0

Cover: Zivi Design, Tel-Aviv, Israel

Print: Febodruk BV, Enschede, The Netherlands

In Silico Strategies For Proteomics Closure

Prediction of small human genes and analysis of high-throughput protein synthesis

In silico benaderingen voor proteomics
Het voorspellen van kleine genen in het humane genoom en
een analyse van high-throughput eiwit synthese

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. W. H. Gispen, ingevolge
het besluit van het college voor promoties in het openbaar te
verdedigen op donderdag 6 april 2006 des ochtends te 10.30 uur

door

Yair Benita

geboren op 18 mei 1973, te Jerusalem, Israel

Promotors:	Prof. dr. D.J.A. Crommelin
	Prof. dr. B. Olivier
	Prof. dr. I. Humphery-Smith
Co-promotors:	Dr. R.S. Oosting
	Dr. M.J. Wise

Dedicated to my parents and Einat for their constant love and support

and to Maytal and Gomzi my endless source of energy and love

TABLE OF CONTENTS

1

General Introduction

The absence of appropriate closure strategies for genomic DNA sequencing initiatives held back that field for many years, namely between 1977 and 1995. During this intervening period many groups were able to rapidly acquire sequence information on the greater part of their target sequences, however, they lacked appropriate 'gap filling' closure strategies. In the past ten year the genomics field has been revolutionized by the ability to use high-throughput technology and transformed radically biological and medical research. Not only do we have the full genomic sequence of hundreds of species, but we are also able to monitor the mRNA expression of thousands of genes simultaneously in cells and tissues.

Proteomics should learn from the past experiences of genomics whenever possible. The term proteome is employed to describe the complete set of proteins encoded by the genome. The study of the proteome, called proteomics, attempts to investigate protein structure, function, protein isoforms, modifications and interactions with other proteins. Proteins carry out most biological functions and knowing which proteins are present, how they interact with each other and what they do is essential to understanding how cells and eventually entire organisms work and behave. Studying the protein content of a tissue, their activity and change during development or disease is essential to our understanding of system level cellular behavior, so as to provide a more holistic and realistic view of biology.

Aebersold separated genomic-style biology into two distinct phases: "a discovery phase to characterize the universe and a browsing phase, in which system-wide biological assays navigate the universe" (Aebersold 2003). The discovery phase of proteomics has seen significant improvements in the identification of genes and their products from the analysis of fully sequenced genomes. The field has also seen improvements in functional protein microarray technologies (Bertone 2005), However, there is currently no experimental platform to systematically measure the diverse properties of proteins at high-throughput. Furthermore, there is no effective technology to systematically and routinely search the proteome. The browsing phase of proteomics is still in its infancy and is mostly characterized by the use of antibody arrays to determine the protein content of a biological sample (reviewed in Pavlickova 2004). The lack of an effective high-throughput technology for proteomics is best demonstrated by the lack of a coordinated total Human

Proteome Project. The reasons for this arise from the combination of the daunting size and complexity of the human proteome, the low abundance of most proteins in tissue and the lack of a PCR equivalent. All high-throughput proteomics technologies, currently available or under development, rely on obtaining a reasonable amount of pure protein product. This is currently one of the rate-limiting steps of high-throughput proteomics. It can be achieved by either isolating proteins from tissues or producing them in suitable expression systems. In human cells and tissues, 10% of the genes have been suggested to encode more than 90% of the protein content (Miklos 2001), while in serum, just four proteins make up more than 90% of the protein bulk. It is therefore inconceivable to obtain large quantities of most proteins from biological samples. Synthetic production of proteins is the method of choice for obtaining large quantities of pure product and is a well established procedure, most commonly performed in *E. Coli* (Makrides 1996; Swartz 1996; Hannig 1998; Baneyx 1999; Baneyx 2004). However, it is inherently difficult to implement on a high-throughput platform. Proteins, unlike DNA, are each different from another with respect to their physicochemical properties. Depending on their amino acids sequence and three dimensional structure, proteins vary in solubility, aggregation, flexibility and other attributes; making it impossible to produce all proteins under similar conditions.

Several approaches have been suggested for the human proteome project (Aebersold 2003; Humphery-Smith 2004). Humphery-Smith suggested to call upon affinity ligands, generated against each open reading frame (ORF) in the human genome, rendering all aspects of proteomics more efficient (Humphery-Smith 2004). This approach requires the synthesis of proteins or protein fragments for generation and selection of appropriate affinity ligands, such as antibodies. Agaton et al. demonstrated the feasibility of generation and screening of affinity ligands on a genomic scale (Agaton 2003). The ability to produce all target proteins or protein fragments in the human proteome is critical for the production and selection of affinity ligands. It relies both on the quality of human genome/proteome annotations for identifying correctly all human proteins and the technical ability to actually produce the target sequences.

HUMAN GENOME ANNOTATION

Genome annotation is the process that identifies sequence features on genomic DNA such as, promoters, known genes, predicted genes and gene models. Current genome annotation with respect to gene finding, is a combination of three approaches (Collins 2003; HGSC 2004): (i) gene prediction programs, (ii) mapping known expressed sequence tags (ESTs) and proven genes (protein-coding, mRNA) onto genomic sequences and (iii) identification of conserved sequences from other vertebrates genome that are likely to be genes. Using this annotation strategy the human genome sequencing consortium estimated the total number of protein-coding genes in the human genome to be in the range of 20,000-25,000 genes (HGSC 2004). The function of most of these genes is still unknown and is most likely to be discovered by the manifested function of the translated proteins (Zhu 2001). Therefore, it is essential for the genome annotation to be of high quality in order for proteomics to complete the next step of functional annotation. The range of 5,000 human genes (20,000-25,000) is attributed mostly to: (i) genes with very small ORFs of less than 100 amino acids; (ii) single exon genes; and (iii) genes that have evolved rapidly. These classes of genes form a gap in the genomics annotation procedure and have to be identified and characterized to complete the human proteome repertoire.

HIGH-THROUGHPUT PROTEIN PRODUCTION

The goal of our group was to make antibody arrays for the quantitative and qualitative analysis of proteins in biological samples. These antibody arrays had to contain 1000’s to 10,000’s of different antibodies, each with known selectivity and cross-reactivity. Antibody arrays will make the study of near-to-total proteome in a high-throughput manner possible and lay the foundations of the human proteome. The first necessary step to make antibody arrays is cloning, expression and purification of the target proteins. Several groups have previously attempted up-scaling of protein production (Christendat 2000; Pizza 2000; Braun 2002; Agaton 2003; Luan 2004; Dobrovetsky 2005). Protein expression success rates were in the range of 50%-80% under denaturing conditions and much lower under non-denaturing conditions (15%-50%). It is inconceivable to express all proteins under the same protocol, therefore, the *a priori* identification of proteins that are suitable for a specific protocol would be a significant achievement that will greatly reduce the already large financial burden. Several groups have previously attempted to link primary protein sequence to its propensity to be expressed in a soluble form rather than in inclusion bodies (Bertone 2001; Goh 2003; Idicula-Thomas 2005; Shimada 2005). Those studies were motivated by low success rates of ~10% in protein expression that were suitable for structural studies. Most of these studies had limited success and in general were unable to suggest a reliable algorithm for prediction of protein expression suitable for structural studies. Prediction of successful protein production has been largely disregarded and a success rate of ~60% is generally considered acceptable in high-throughput projects.

AIM AND OUTLINE OF THE THESIS

This thesis will focus on strategies to improve human genome annotations with respect to gene finding and high-throughput protein production pushing proteomics one step further towards closure.

The aims of this thesis are:

1. To identify new small human genes, encoding proteins smaller than 100 amino acids.
2. To establish a platform for efficient high-throughput protein production.

The thesis is divided into two main parts. In the first part a study into the properties and identification of small human genes is presented. This class of genes encodes proteins smaller than 100 amino acids and is considered problematic for both genomics and proteomics. A thorough overview of the challenges in predicting those genes computationally and identifying them biologically is given in chapter 2; together with an analysis of currently annotated small human genes, their genomic structure and attributes. Despite many advances in eukaryotic gene finding, this class of genes has been largely overlooked. There is a strong bias in current genome annotation strategies towards long, highly expressed and conserved genes. Therefore, in chapter 3, a novel method for identifying new small human genes is outlined. This method is based on initial *in silico* prediction of small genes that lays the foundation for gene specific identification from biological samples.

In part two of this thesis a thorough analysis of a protein production pipeline is presented. The high-throughput pipeline and initial success rate observations are described in chapter 4. In this chapter two main ‘gaps’ have been identified where the success rate significantly dropped, namely, PCR and protein expression. Chapter 5 outlines the computational analysis of 1,438 human exons and a method to computationally identify DNA sequences that are not suitable for the PCR protocol used. The next step of expression and purification of human proteins is discussed in chapter 6. Hundreds of proteins (617) were characterized from sequence alone and correlated to successful expression and expression level. A second study that investigated the correlation between inclusion body formation and successful protein expression is presented.

REFERENCES

Aebersold R. 2003. Constellations in a cellular universe. *Nature* **422:** 115-6.

Agaton C., Galli J., Höidén Guthenberg I., Janzon L., Hansson M., Asplund A., Brundell E., Lindberg S., Ruthberg I., et al. 2003. Affinity proteomics for systematic protein profiling of chromosome 21 gene products in human tissues. *Mol Cell Proteomics* **2:** 405-14.

Baneyx F. 1999. Recombinant protein expression in *Escherichia coli*. *Curr Opin Biotechnol* **10:** 411-21.

Baneyx F. and Mujacic M. 2004. Recombinant protein folding and misfolding in *Escherichia coli*. *Nat Biotechnol* **22:** 1399-408.

Bertone P. and Snyder M. 2005. Advances in functional protein microarray technology. *FEBS J* **272:** 5400-11.

Bertone P., Kluger Y., Lan N., Zheng D., Christendat D., Yee A., Edwards A.M., Arrowsmith C.H., Montelione G.T. and Gerstein M. 2001. SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res* **29:** 2884-98.

Braun P., Hu Y., Shen B., Halleck A., Koundinya M., Harlow E. and LaBaer J. 2002. Proteome-scale purification of human proteins from bacteria. *Proc Natl Acad Sci U S A* **99:** 2654-9.

Christendat D., Yee A., Dharamsi A., Kluger Y., Gerstein M., Arrowsmith C.H. and Edwards A.M. 2000. Structural proteomics: prospects for high throughput sample preparation. *Prog Biophys Mol Biol* **73:** 339-45.

Collins J.E., Goward M.E., Cole C.G., Smink L.J., Huckle E.J., Knowles S., Bye J.M., Beare D.M. and Dunham I. 2003. Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Res* **13:** 27-36.

Dobrovetsky E., Lu M.L., Andorn-Broza R., Khutoreskaya G., Bray J.E., Savchenko A., Arrowsmith C.H., Edwards A.M. and Koth C.M. 2005. High-throughput production of prokaryotic membrane proteins. *J Struct Funct Genomics* **6:** 33-50.

Goh C.S., Lan N., Echols N., Douglas S.M., Milburn D., Bertone P., Xiao R., Ma L.C., Zheng D., et al. 2003. SPINE 2: a system for collaborative structural proteomics within a federated database framework. *Nucleic Acids Res* **31:** 2833-8.

Hannig G. and Makrides S.C. 1998. Strategies for optimizing heterologous protein expression in *Escherichia coli*. *Trends Biotechnol* **16:** 54-60.

HGSC 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931-45.

Humphery-Smith I. 2004. A human proteome project with a beginning and an end. *Proteomics* **4:** 2519-21.

Idicula-Thomas S. and Balaji P.V. 2005. Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in *Escherichia coli*. *Protein Sci* **14:** 582-92.

Luan C.H., Qiu S., Finley J.B., Carson M., Gray R.J., Huang W., Johnson D., Tsao J., Reboul J., et al. 2004. High-throughput expression of *C. elegans* proteins. *Genome Res* **14:** 2102-10.

Makrides S.C. 1996. Strategies for achieving high-level expression of genes in *Escherichia coli*. *Microbiol Rev* **60:** 512-38.

Miklos G.L. and Maleszka R. 2001. Protein functions and biological contexts. *Proteomics* **1:** 169-78.

Pavlickova P., Schneider E.M. and Hug H. 2004. Advances in recombinant antibody microarrays. *Clin Chim Acta* **343:** 17-35.

Pizza M., Scarlato V., Masignani V., Giuliani M.M., Aricò B., Comanducci M., Jennings G.T., Baldi L., Bartolini E., et al. 2000. Identification of vaccine candidates against serogroup B *meningococcus* by whole-genome sequencing. *Science* **287:** 1816-20.

Shimada K., Nagano M., Kawai M. and Koga H. 2005. Influences of amino acid features of glutathione S-transferase fusion proteins on their solubility. *Proteomics* **5:** 3859-63.

Swartz R.S. 1996. *Escherichia coli* recombinant DNA technology, In *Escherichia coli and Salmonella* (ed. Neidhardt), pp. 1693-711. ASM, Washington DC.

Zhu H. and Snyder M. 2001. Protein arrays and microarrays. *Curr Opin Chem Biol* **5:** 40-5.

Part I

ADVANCING TOWARDS COMPLETE GENOME ANNOTATION

Confronting small protein-coding human genes

2

Challenges in the annotation of small genes in the human genome

Yair Benita[1], Ronald S Oosting[1], Michael J Wise[2], Zhenyu Xuan[3], Michael Q Zhang[3] and Ian Humphery-Smith[4]

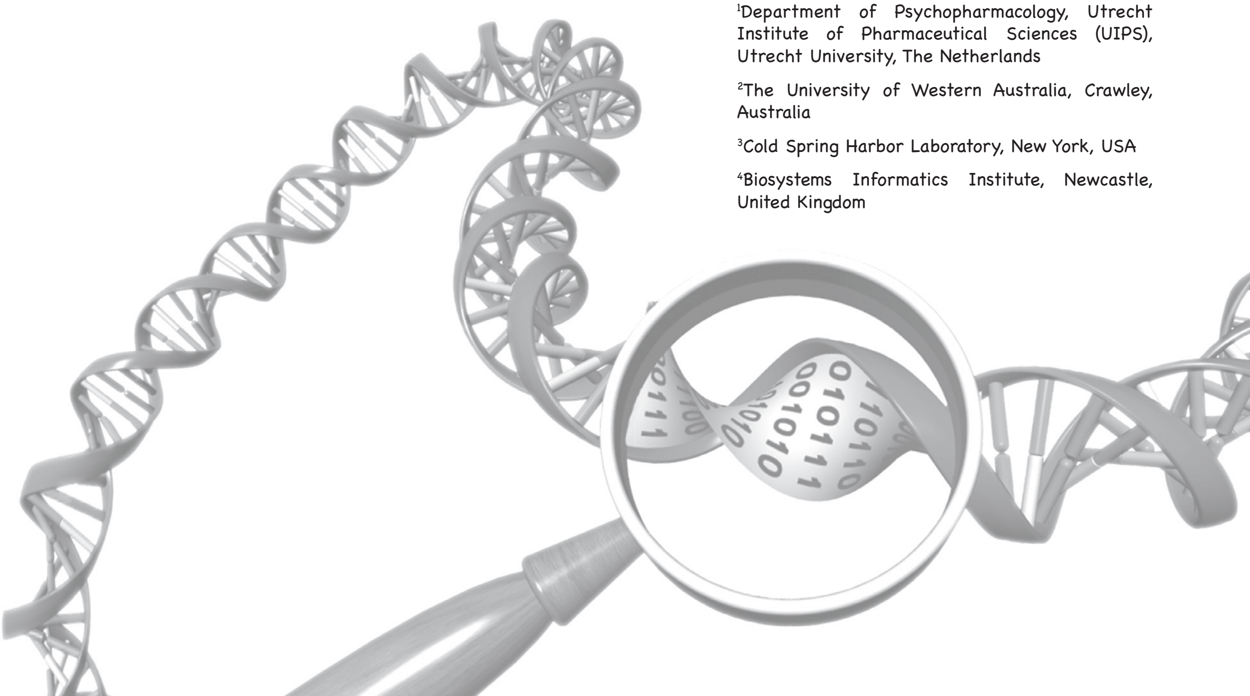
Submitted for publication

[1]Department of Psychopharmacology, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, The Netherlands

[2]The University of Western Australia, Crawley, Australia

[3]Cold Spring Harbor Laboratory, New York, USA

[4]Biosystems Informatics Institute, Newcastle, United Kingdom



ABSTRACT

The detection of small open reading frames (ORF’s) of less than 300 bp remains a significant challenge to ‘completion’ of gene annotation within the human genome. The smallest known human ORF contains three exons and encodes a 33 amino acid protein. Small annotated ORF’s were shown to possess a gene structure similar to other human genes, including similarity of exon structure. Surprisingly, very few small genes were composed of a single exon. On average they contained two or three exons and as many as seven, the smallest being only 3 bp and detectable only experimentally. The performance of several extant gene-finding software algorithms was assessed for this most difficult of gene classes with respect to a known small gene set of 228 ORF’s. Those that scored best did so due to appropriate weighting of evolutionary conservation.

1. INTRODUCTION

Recently, the international human genome sequencing consortium announced the completion of the euchromatic sequence of the human genome (HGSC, 2004). The total number of protein-coding genes in the human genome is estimated to be in the range of 20,000-25,000. This range of 5,000 genes is attributed mostly to: (i) genes with very short open reading frames (ORFs) of less than 100 amino acids; (ii) single exon genes; and (iii) genes that have evolved rapidly.

This review will focus on small genes in the human genome. Studies that specifically targeted this class of genes in the human genome and other organisms will be presented, followed by a detailed overview of currently annotated small genes in the human genome. Here, the term “ORF” is employed to refer to a continuous sequence in DNA or RNA that can potentially encode a protein or part of a protein. The term “gene” to refer to a contiguous or non-contiguous DNA sequence that actually encodes a protein and the term “small” to refer to genes and proteins comprised of less than 100 codons or amino acids, respectively.

Small proteins found in yeast include a number of important classes, such as mating pheromones, energy metabolism, proteolipids, chaperonins, stress proteins, thioredoxins and metal ion chelators (Basrai et al. 1997). Others, found in *Escherichia coli* (*E. coli*) include DNA and RNA binding proteins and ribosomal proteins (Rudd et al. 1998). This review will focus on small proteins encoded by small genes as opposed to small functional proteins which are the products of cleavage. Examples of cleavage products from the human proteome include insulin, a peptide cleaved from a 110 amino acids (aa) precursor (Bell et al. 1980) and POMC, a 267 aa protein which is cleaved into several smaller neuropeptides (Takahashi et al. 1981).

A threshold of 100 codons for an ORF was used in the annotation of the yeast genome when the entire sequence of *Saccharomyces cerevisiae* chromosome III was announced (Oliver et al. 1992). As the yeast genome contains mostly continuous non-spliced ORFs, it was natural for a threshold to be placed above that which an ORF is likely to encode a protein. The 100 codon threshold was rationalized by suggesting that 3 of the 64 codons in the genetic code are termination codons, therefore the likelihood of not having a termination codon in a random continuous DNA sequence of 100 codons is less than 5% (Fickett 1994; Fickett 1995). As such, this threshold was initially accepted as an arbitrary compromise so as not to miss the majority of protein-coding genes and to avoid a large number of non-coding ORFs. It allowed for a good signal/noise ratio i.e. a ratio of functional genes to all possible ORFs, which in the case of yeast was ~64% (4,800/7,472) for the 100 codons threshold and ~13% (5,400/41,005) for 20 codons. As the threshold is lowered the signal/noise ratio drops exponentially, making it increasingly harder to distinguish real functional genes from fortuitous ORFs. Although the 100 codon rationale was based on an assumption of randomness, DNA sequences are far from random. Examination of the entire yeast genome by Mackiewicz et al demonstrated the existence of 7,472 ORFs longer than 100 codons and concluded that only 4,700-4,800 were protein coding genes (Mackiewicz et al. 1999). Their conclusion was based on MIPS annotations (Mewes et al. 2004), homology to known proteins, and the observation that many ORFs had properties of antisense sequences of coding ORFs. A similar analysis examined all 41,005 ORFs longer than 20 codons and suggested that the genome contained 5,300-5,400 protein-coding genes (Mackiewicz et al. 2002). Termier and Kalogeropoulos analyzed yeast ORFs using discriminant analysis that combined ORF length, codon bias, di-peptide and amino acid composition. They showed that some ORFs longer than 100 codons were likely to be fortuitous, suggesting that the 100 codons threshold rule was further inappropriate. The Codon Adaptation Index (CAI) was defined by Sharp and Li based on the observation that an organism had a codon bias in the sense that highly expressed genes were preferentially encoded. For instance, the codons AGA and AGG, which encode the amino acid arginine, are almost never used in highly- expressed genes of *E. coli* (Jansen et al. 2003; Sharp and Cowe 1991). Such codon bias was successfully used to identify coding ORFs below the 100 codons threshold (Barry et al. 1996; Yada and Hirosawa 1996) and was also shown to be applicable to the yeast genome (Jansen et al. 2003), thereby making it easier to predict genes correctly. Furthermore, Cebrat et al. showed that bias existed even within a codon at the first, second and third nucleotides. They devised a method to graphically display any bias at gene and genome levels, enabling them not only to distinguish a coding sequence from an intergenic sequence, but also to identify correctly the strand of the ORF and the phase of the coding sequence (Cebrat et al. 1998). By 1998 gene prediction in yeast has evolved significantly, however, the 100 codon threshold was still considered a useful safety margin by many, since a long predicted gene still had a higher probability of being correct than a short predicted gene.

In recent years, annotation efforts shifted towards the human genome as the full genomic sequence became available. While it was relatively easy to analyze all possible ORFs above a threshold in bacteria and yeast, the human genome was much more challenging in this respect. Not only were genes alternately spliced, but the genome was also much larger with less than 1% protein-coding sequences (HGSC 2004). ORF length and codon usage were no longer good gene predictors by themselves and gene prediction became much harder. Several gene prediction programs were developed for mammalian genomes (Burge and Karlin 1997; Korf et al. 2001; Parra et al. 2003; Parra et al. 2000; Solovyev 2001). Most of these programs were trained using known annotated genes and were expected to identify new genes with similar structure and properties. Despite significant improvements in mammalian gene prediction, small genes remain a challenge.

The exponential growth of available biological data with respect to gene transcription (expressed sequence tags (ESTs) and full cDNA sequencing projects) pushed forward interest in genome annotation (Schuler 1997; Schuler et al. 1996). The increased number of ESTs and full length cDNA in revision 3 of the *Drosophila melanogaster* genome annotation altered 85% of the gene models annotated previously in revision 2 and added 273 novel genes (Misra et al. 2002). The annotators then set the acceptable ORF threshold at a minimum of 50 codons, but most of those were not annotated as genes without supporting experimental data. In their second revision of human chromosome 22, despite the accumulating amount of experimental data, Collins et al still used the 100 codons threshold as the detectable limit for a protein-coding genes (Collins et al. 2003).

2. THE CHALLENGE OF IDENTIFYING SMALL GENES AND THEIR PRODUCTS

In this section the challenge of predicting and identifying small genes at mRNA and protein levels is discussed. Where available, studies that targeted small genes specifically are presented.

2.1 The challenge of predicting small genes *in silico*

The two most important aspects of any gene prediction program are the type of information used and the algorithm employed to combine that information. Three types of information are used in gene prediction: DNA signals, such as splice sites; content statistics, such as codon bias; and similarity to known genes and proteins (Stormo 2000). Algorithms employed by most gene prediction programs can be classified into two categories: dynamic programming systems that identify and score exons based on signals and content statistics, combining these into the highest scoring gene; and Hidden Markov Models (HMMs) in which a gene model is defined using several states (exon, intron, etc.) and associated probabilities to assign transitions between states. When an HMM is in a particular state it “emits” DNA sequence characteristics for that state (for a review see Durbin et al. 1998). Examples of the former include Morgan (Salzberg et al. 1998) and Fgenes (Solovyev et al. 1995) and examples of the later include Genscan (Burge and Karlin 1997), Genie (Kulp et al. 1996) and HMM-gene (Krogh 1997).

The small number of codons and amino acids available for statistical analysis in small genes limits the efficiency of gene prediction programs. While small genes may have DNA signals like any other gene, parameters related to content statistics may not reach significance level. Programs targeting specifically short ORFs were mostly developed for bacterial and yeast genomes, where the majority of genes are contiguous. GeneHacker, an HMM gene prediction program developed for microbial genomes was applied to the *Cyanobacterium* and detected 7 of the 8 annotated small genes, 36 very short ORFs (< 50 bp) of which 4 had homologs in other databases; and 57 moderately short ORFs (150-300 bp) of which 5 had homologs (Yada and Hirosawa 1996). Barry et al. applied a discriminant function, using in-phase hexamer frequencies, to identify short ORFs of 36-100 codons in yeast and identified 52/58 annotated genes (Barry et al. 1996). Andarde et al. identified 10 new short ORFs in the yeast genome based on codon usage, amino acid composition and homology to other organisms (Andarde 1997).

Mammalian genomes contain mostly non-contiguous genes with the majority of coding exons in the range of 100 to 200 bp. Therefore, gene prediction in mammalian genomes can be considered akin to the prediction of small coding sequences, particularly since programs essentially predict exons which are later combined into a gene structure. Some programs focus on predicting only exons, such as MZEF (Zhang 1997) and Exoniphy (Siepel and Haussler 2004). MZEF was designed to predict only internal coding exons in genomic DNA. It combines hexamer frequency with 9 variables, such as, exon length and exon-intron transitions and uses quadratic discriminant analysis to detect exons. Exoniphy employs phylogenetic HMMs to predict evolutionarily conserved exons. Both programs have demonstrated the ability to predict short coding ORFs (Rogic et al. 2001; Siepel and Haussler 2004). Recently, Gao and Zhang compared 19 different algorithms for identifying short coding human sequences based on content measures alone (Gao and Zhang 2004). Prediction accuracy for exons in the range of 42 to 192 bp was estimated. Current algorithms were able to predict coding sequences of 42 bp, 63 bp, 87 bp, 108 bp, 129 bp, 162 bp and 192 bp at best with an accuracy of 83.4%, 87.2%, 89.9%, 91.9%, 93.4%, 95.1% and 96.2%, respectively. These results demonstrate the improvements achieved in the field of gene prediction since the first algorithms applied to yeast. More than 90% of the nucleotides can now be correctly identified as either coding or non-coding (Stormo 2000).

However, identifying exons correctly is only one part the prediction process. Those exons must be combined correctly into genes, a task for which accuracy drops significantly even for the best gene prediction programs (Rogic et al. 2001; Stormo 2000; Zhang and Zhang 2002). An exon must be examined in the context of the gene and the assembly of exons into a gene is often used to enhance or reduce an exon's likelihood of being correct. For instance, an ORF must be maintained through neighboring exons and a multi-exon gene must be made of a first exon, internal exons and a terminal exon. Those restrictions reduce the number of possibilities when combining exons into a gene. However, if a single exon is missed or is incomplete, the entire prediction will be incorrect.

HMMGene, an HMM gene finder was applied to the well annotated *Adh* region of the *Drosophila melanogaster* using EST database matches (Krogh 2000). It was shown that due to the high noise in EST data, the specificity decreased and the number of incorrect predictions increased. Artifacts in the EST data, or ESTs that match the wrong strand caused HMMGene to incorrectly predict "silly" small genes. The author concluded that including larger ESTs, above a critical length (specific length not mentioned), could be considered "safe" and beneficial. However, ESTs were not easy to deal with in an automated manner and more work is needed to fully benefit from them.

Most of human gene prediction programs do not implement a minimum threshold for exon or gene sizes. The MZEF algorithm, for instance, implements exon length as a discriminating variable but employed no specified threshold for defining an ORF. When MZEF was applied to human chromosome 22 (UCSC build HG16), internal exons as small as 18 bp were predicted. Similarly, Genscan, incorrectly predicted a one–nucleotide-long exon (genscan id: NT_011109.592) in human chromosome 19 (Burge and Karlin 1997), while the smallest gene predicted by Genscan in the human genome was a two exons gene with a coding sequence of 27 bp (genscan id: NT_007933.94). Furthermore, almost 12% of the genes predicted in the human genome (build HG17) by Genscan had a coding sequence smaller than 300 bp. Despite having no threshold for defining genes and/or exons, most gene prediction programs were trained using known annotated genes in order to identify new genes similar to those in the training set. In general, when putting together a training set, scientists apply strict rules to the selection of sequences used in the training set, in order to establish a high quality set with as few mistakes as possible. However, in most cases, this leads to sets in which small genes are likely to be under-represented or non-existent (few small genes are currently annotated, see next section). Thus, most programs are likely to be biased against finding small genes or are less likely to accurately predict them compared to longer ones.

Predicting all coding exons correctly and combining them into genes represents a major challenge to genome annotators. Prediction of small genes in the human genome presents an even greater one. Exons of small genes are likely to be on the low end or even smaller than the 100-200 bp exon length range in most genes. As far as we are aware, there is currently no gene prediction program that targets specifically small genes in the human genome.

2.2 Pseudogenes introduce difficulties in identifying small genes

Pseudogenes are gene copies that have lost the ability to code for a protein and are typically identified through annotation of disabled, decayed, duplicated, or incomplete protein-coding sequence. Pseudogenes have been shown to exist in prokaryotes with an occurrence of up to 5% of all gene-like sequences (Liu et al. 2004). In the human genome, there is evidence suggesting that there may be twice as many pseudogenes as protein-coding genes (Gibbs et al. 2004; Torrents et al. 2003; Waterston et al. 2002; Zhang et al. 2003). The high number of pseudogenes complicate the task of predicting and annotating functional protein-coding genes, especially intronless genes (Zhang 2002).

Sequence based methods for identifying pseudogenes rely on the presence of frame shifts or internal stop codons. However, as many as half of known pseudogenes do not contain these elements (Torrents et al. 2003). Such pseudogenes can only be distinguished from the real protein-coding genes, if the genes are not fast evolving, by estimating the ratio of the rates of substitution at synonymous sites to the rate of substitution at non-synonymous sites; or by looking at the pattern of substitution in conserved protein domains (Balakirev and Ayala 2003; Coin and Durbin 2004; Torrents et al. 2003). The ability to distinguish such pseudogenes from real genes decreases as coding sequences become smaller.

Unprocessed pseudogenes are a class of pseudogenes that arise from tandem duplication or unequal crossing-over and retain the exon-intron structure of the parental gene (Zhang and Gerstein 2004). Their gene structure is often incomplete and includes only two or three exons. Small genes, which are likely to have few exons, are more likely to be mistaken as pseudogenes and the sheer number of pseudogenes in the human genome introduces further difficulties in the identification and annotation of small genes.

2.3 Small genes and non- coding RNAs

Non-coding RNAs (ncRNAs) are all RNAs other than mRNA, including, microRNA, siRNA, snRNA, rRNA and others. This class of genes, despite their biological importance, has gone relatively undetected for a long period. Current annotation strategies for protein-coding genes, essentially do not work at all for non-coding genes (Eddy 2001). This class of genes should be distinguishable relatively easily even from small protein-coding genes. The most obvious attributes, when comparing ncRNAs to small protein-coding genes, are the lack of an ORF; and their characteristic secondary structure.

2.4 The challenge of detecting mRNAs of small genes

ESTs and full-length cDNA projects have been valuable for the annotation of eukaryotic genomes. However, the sampling of mRNA from tissue is a random process and the selection of cDNA fragments above a certain size to be sequenced makes those cDNA libraries biased towards longer highly-expressed genes (Sambrook and Russell 2001). The likelihood of identifying small genes by random sequencing of mRNA is low. However, mRNA of small genes may be identified if specifically targeted as discussed below.

Olivas et al. used an experimental approach to locate previously undetected small ORFs in yeast. Based on the observation that the yeast genome has a very compact distribution of genes, genomic sequences larger than 2kb were targeted, in which no known or predicted ORFs were reported. Using PCR to amplify 58 such regions followed by northern blot of total yeast RNA, 15 new RNA transcripts ranging in size from 450 to 1200 bp were found. Two of these appeared to be non-coding RNAs while the others had ORFs in the range of 84 to 98 codons. These mRNA had characteristics akin to their larger cousins, such as poly(A) tails and a relatively long 3′ untranslated region. Reymond et al. 2002 combined data from gene prediction programs and predicted CpG islands with mapping of ESTs containing bona fide 3′ ends to create a reliable set of new potential genes in the human genome. Twenty seven putative genes were identified in chromosome 21, of which 19 were subsequently validated by 5′ and 3′ RACE. Most of these transcripts were small and were suggested to encoded small proteins. The smallest one was a three exon gene encoding a 33 amino acid protein of unknown function (Genbank id: *AF426268*).

2.5 The challenge of isolating small proteins

To study and characterize proteins from a crude mixture of cell extract, the proteins must be separated and detected. Separation techniques include 2D gel electrophoresis, affinity chromatography to selectively capture proteins and size exclusion chromatography. Detection methods include Edman sequencing and mass spectrometry (for a review on separation and detection methods see Rabilloud 2000). Identification of small proteins is difficult as these proteins have a low molecular mass and are often present at low abundance, requiring highly sensitive methods (Wasinger et al. 1995). Furthermore, it is difficult to discriminate a real small protein from a breakdown product of larger proteins.

The classical method of 2D gel electrophoresis has been used for decades to study proteins. However, resolution of 2D gels is relatively low and Edman sequencing, used for identification of the protein, requires large amounts of purified protein, making this method unsuitable for low molecular weight proteins. Yet, in one study, Wasinger and Humphery-Smith were able to resolve 42 low molecular weight *E. coli* proteins from 2D gels followed by Edman microsequencing. Fourteen ORFs were confirmed and corresponded to small proteins. Nevertheless, the authors mentioned that small proteins were inherently difficult to study due to their potentially low intracellular concentration, poor solubility, high diffusibility during electrophoresis and/or column chromatography and the reduced number of amino groups and dye/isotope accepting elements (Wasinger and Humphery-Smith 1998).

To be applied successfully to low molecular mass proteins at a proteomic level, methods must couple adequate separation and detection. Detection of proteins has been greatly facilitated by recent advances in biological mass spectrometry, which has become the method of choice for protein characterization (Mann et al. 2001). Proteins excised from gels or in complex mixture solutions, can be digested by proteases and the peptides can be sequenced by tandem mass spectrometry. Proteomics methods, which couple separation and detection, have been applied on a large scale by first reducing crude protein mixtures to peptides, applying 1D or 2D chromatography followed by mass spectrometry (Gygi et al. 2000; Martin et al. 2000; Yates et al. 1997). Those methods were shown to be useful for characterizing the medium to high molecular mass proteome (20-200 kDa), but difficult to apply to the lower molecular mass spectrum. On the other hand, Wang et al. (2002) compared three separation and detection schemes and isolated *E. coli* components in the range of 2 to 20 kDa. This study showed the tractability of separating low molecular mass proteins by HPLC followed by MALDI-TOF mass spectrometry detection. Although some components were common to the three detection schemes employed, other unique components were found for each methodology, emphasizing both the bias and the complementarity of each approach. The resolving power of complex protein mixtures using HPLC is increased linearly by the amount of starting material. To better study the low molecular mass proteome, Keller et al. devised a method for concentrating proteins from individual HPLC fractions, while another approach suggested that many small proteins are secreted and therefore protein mixtures may be enriched for small proteins from supernatants of cell lines or obtained from body fluids such as serum or cerebrospinal fluid so as to allow the elucidation of small ORFs (Mann et al. 2001).

Once peptides resolved from a protein mixture have been identified, they could be used to search genomic databases to identify coding regions (Pandey and Lewitter 1999). As such, peptide sequences can help establish real exons. However, it is very difficult to achieve total coverage of a protein by mass spectrometry at the low levels usually available in biological samples (Mann et al. 2001). The identification of the low molecular mass proteome is an on-going research effort facing many challenges. Despite significant improvements in proteomics research and the ability to better characterize proteins, the application of proteome data to genome annotation with respect to gene finding has not been widely adopted by genome annotators.

3. SMALL ANNOTATED GENES IN THE HUMAN GENOME

In this section we present small annotated human proteins that are encoded by small genes, their function and structure. The data presented here can be viewed graphically using a gbrowse (Stein et al. 2002) based genome browser at http://rulai.cshl.edu/smallgenes.

3.1 Genomic structure of small genes

Human genes that were well annotated at the genome and proteome levels had a CDS size in the range of 126 bp to 26,394 bp with an average of 1,556 bp (Figure 1). Less than 3% of these genes (228 genes) were smaller than 300 bp with an average of 244 bp.

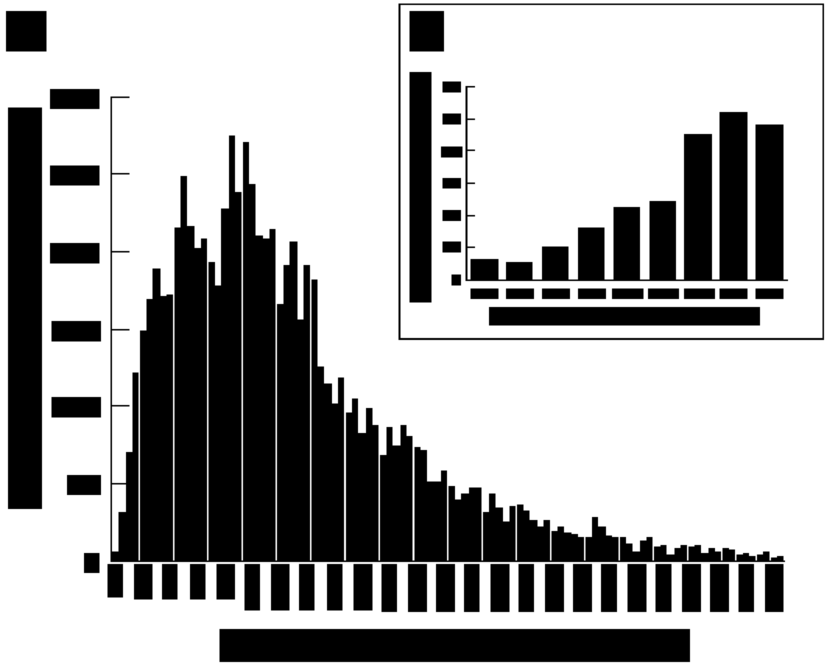


Figure 1: Distribution of coding sequence length of 8,235 coding sequences from the NCBI CCDS database (http://www.ncbi.nlm.nih.gov/CCDS) up to 5,000 bp (A) and up to 300 bp (B). The selected CDS were obtained by comparing Swissprot (Bairoch et al. 2005) protein sequences to CCDS sequences. Those that matched with an identity and sequence coverage of at least 95% were used in this analysis. Column width for A and B is 50 bp and 20 bp, respectively, and the length label represents the midpoint of the column.

Most annotated small human genes had 2 or 3 coding exons and, surprisingly, as many as 7 exons (Figure 2). The two small genes with 7 coding exons were 270 bp and 264 bp long respectively. Both were precursors with *FXYD* domains and were annotated as involved in ion transport (Figure 3). As small genes have a relatively small coding sequence, we expected to observe a relatively high number of single exon genes. However, only 12% of the annotated small human genes were single exon genes compared with 9% for larger genes. Pseudogenes annotated in the Vega database (Vertebrate Genome Annotation Database (Ashurst et al. 2005) had up to 82 exons per gene, with the vast majority of pseudogenes containing up to three exons.

Annotated multi-exon small genes had exons which were, on average, shorter than exons of larger genes (Figure 4). Considering all exons of small genes, the proportion of initial and terminal exons, which may be partially coding, to internal exons, most of which are fully coding, is relatively high and may bias the coding exon size comparison. However, even internal exons of small genes were significantly smaller than those of large genes, with an average of 83 bp and 142 bp ($p < 0.01$), respectively. The smallest gene in our reference dataset of 228 small genes was 41 codons long, containing two coding exons: 34 bp and 92 bp (CCDS ID: *CCDS12746.1*). The smallest coding internal exon detected in our set of 8,235 genes is 3 bp long. The smallest coding internal exon in our set of 228 small genes is 12 bp long. Here, only internal exons were considered, because the initial and terminal exons are often only partially coding. The smallest protein entry in the SwissProt database is 3 aa long (human growth modulating peptide GRWM_HUMAN), However, this is most probably a cleavage product derived from a larger protein. To the best of our knowledge, the smallest annotated human gene that encodes a protein was found by Reymond et al. 2002 (Reymond et al. 2002). The authors employed a combination of computational and experimental approaches to detect additional ORFs in chromosome 21 and found a three exon gene with a 33 codons long ORF. The smallest known functional gene-product found in *E. coli* is a 29-amino acid peptide involved in K^+ transport (KdpF) (Koonin et al. 1997). ORFs with as few as 14 amino acids have been predicted in *E. coli* and 28 amino acids in *S. cerevisiae* (Barry et al. 1996), while Hendrix (personal communication) detected a highly probable small gene in the N15 bacteriophage and a small gene within the *E. coli* Shiga-like toxin operon of, respectively, 18 and 13 amino acids. More recently,

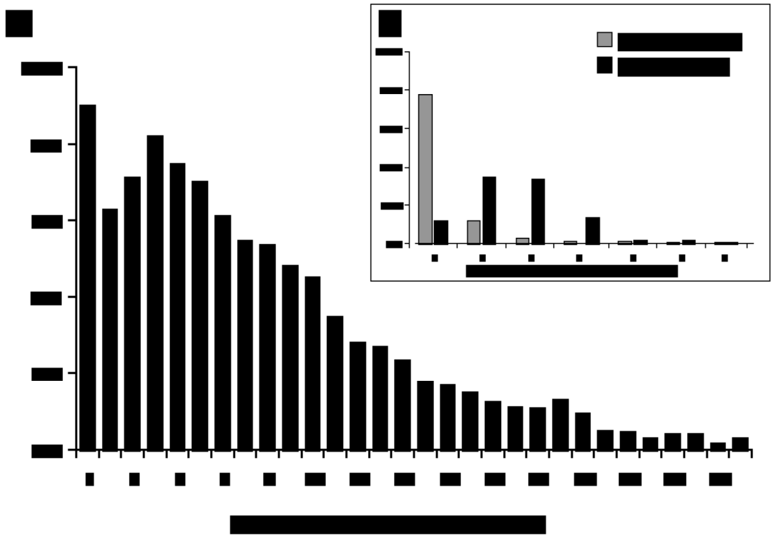


Figure 2: Distribution of the number of coding exons per gene for the 8,235 genes (A) as described in Figure 1, for the subset of 228 genes (B-small genes) with CDS smaller than 300 bp and of 4,450 pseudogenes (B-pseudogenes) across 9 chromosomes (6, 7, 9, 10, 13, 14, 20, 22 and X) annotated in the Vega database (Ashurst et al. 2005). Nine percent of the genes were single exon genes. Multiple exon genes had up to 144 exons per gene with an average of 9.5.

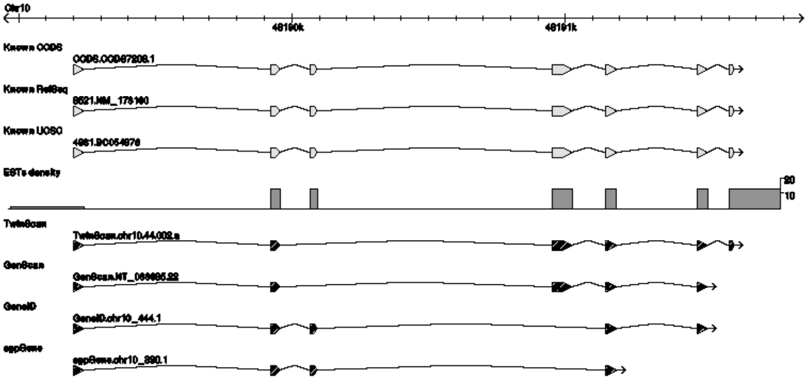


Figure 3: GBrowse generated image of *FXYD4*, a seven exon small gene of 270 bp encoding a precursor protein of an ion transport regulator. The light gray transcripts show known gene annotations from CCDS (http://www.ncbi.nlm.nih.gov/CCDS/), RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq) and UCSC (http://genome.ucsc.edu). ESTs from dbEST (http://www.ncbi.nlm.nih.gov/dbEST) were mapped to the human genome and only ESTs with a coverage of at least 50% and identity of at least 90% to the genome sequenced were clustered to form the density track. Gene prediction data from various sources was used, as distributed by UCSC (Karolchik et al. 2003), and is shown in gray stripes.

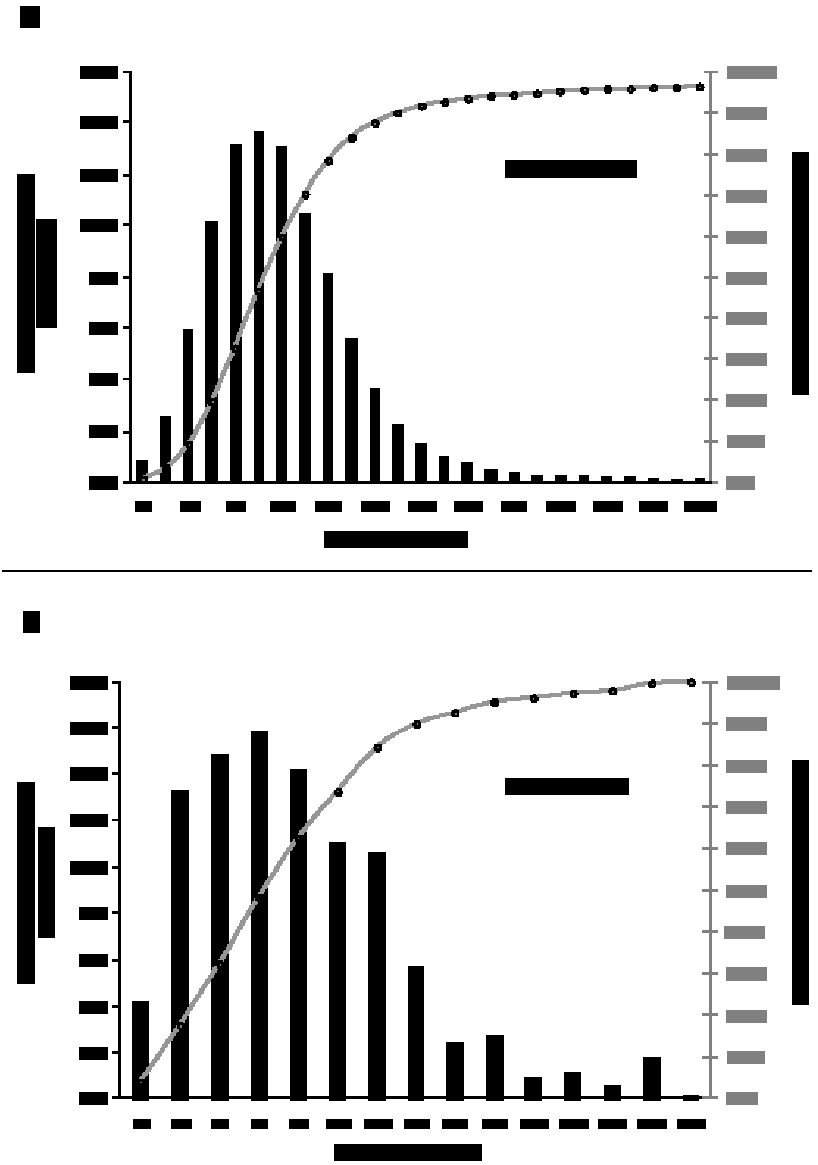


Figure 4: Distribution of coding exon length from large genes (A) and small genes (B) with multiple exons. The selected genes were the same as in Figure 1. The column width used was 20 bp and the category x label represents the midpoint of the column. The gray line shows the cumulative percentage of the total 76,536 exons (A) and 584 exons (B) in the category. The difference in average exon size was found to be highly significant (p<0.01).

artificial constructs encoding just six amino acids were able to be transcribed and resulted in functional gene-products involved in intracellular signalling in *B. subtilis* (Lazazzera et al. 1997). Some of these small gene-products are conceivably 'Leader' peptides (Rudd et al. 1998). Confirmation of such small ORF's will prove extremely difficult statistically and/or experimentally.

Most of the coding sequences of annotated small human genes were shown to be contained within one exon (Figure 5). In the case of small genes with more than two exons, the largest coding exon was an internal exon in 63% of the cases. For example, among the 7 exons of the *FXYD4* gene (Figure 3), the largest exon was 75 bp long and the other six were in the range of 20-40 bp long. Furthermore, the introns of *FXYD4* were also relatively small as the entire CDS stretched over 2,420 bp of genomic DNA.

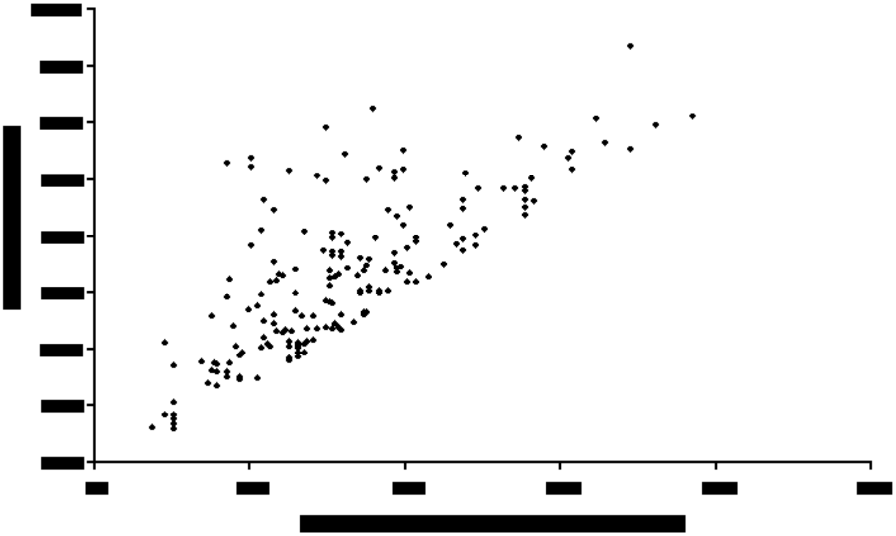


Figure 5: Total CDS coverage as a function of the largest exon for 200 small genes with multiple coding exons, described in Figure 1. The average largest exon was 131 bp long with a coverage of 54%.

3.2 Evaluation of small genes' prediction accuracy

As small genes have a short coding sequence, they are harder to predict than longer genes (see introduction). In several projects that attempted to annotate genomes, assumptions were made that some genes may be missed, especially genes encoding small proteins (Collins et al. 2003; HGSC 2004; Misra et al. 2002). Gene prediction programs were previously evaluated at the nucleotide, exon and protein levels (Burset and Guigo 1996; Guigo et al. 2000; Zhang 1997). However, no specific reference was made to small genes and the ability to predict their occurrence correctly as CDS size became smaller.

Gene prediction programs are tuned by their creators (or by users) to have either a high sensitivity at the expense of missing some real genes or to predict as many genes as possible risking a higher frequency of false positives. Twinscan is an example of the former, predicting ~20,000 genes in the human genome and Genscan is an example of the latter, predicting ~37,000 genes (UCSC annotations, HG16). We employed four existing *ab initio* gene prediction programs that represent a variety of algorithms currently in use. Regardless of the programs tuning and accuracy in predicting human genes, the four programs showed reduced performance in predicting correctly small genes as compared to predicting large genes with respect to a known set of 8,007 large genes (Figure 6). Specificity values at the nucleotide level were very similar between the prediction of small and large genes, suggesting that when nucleotides were predicted as coding they were actually coding with similar confidence in small and large genes. However, sensitivity values were lower for small genes, meaning that the proportion of coding nucleotides that were predicted correctly as coding were lower for small genes. In other words, many coding nucleotides in small genes were missed. That observation was also supported by the higher proportion of missed exons in small genes. The sgpGene prediction program had the smallest difference in prediction between small and large genes. sgpGene uses alignments to known proteins to improve prediction and the small difference in sensitivity between the two groups can be explained by the association of many small genes to small-gene families (see small human proteins section below).

In general, gene prediction programs performed better in identifying single exon large genes than single exon small genes. Genscan, for instance, missed 61% of the small single exon genes and only 6% of the large single exon genes. That observation is consistent with the 100 codons threshold, suggesting that single exon genes are easier to predict as they become longer and are more likely to be real coding genes. In small single exon genes, there is a much higher rate of incorrectly identifying the opposite strand as the coding sequence (up to 7%). Cebrat et al. 1998 showed that one of the properties of the genetic code is that it generates an ORF inside a coding sequence, in a specific phase of the antisense strand, with much higher probability than in random DNA sequence. The antisense ORF was shown to possess the same features as the real gene. This observation could explain the higher rate of errors in predicting the correct strand when the coding sequence is short.

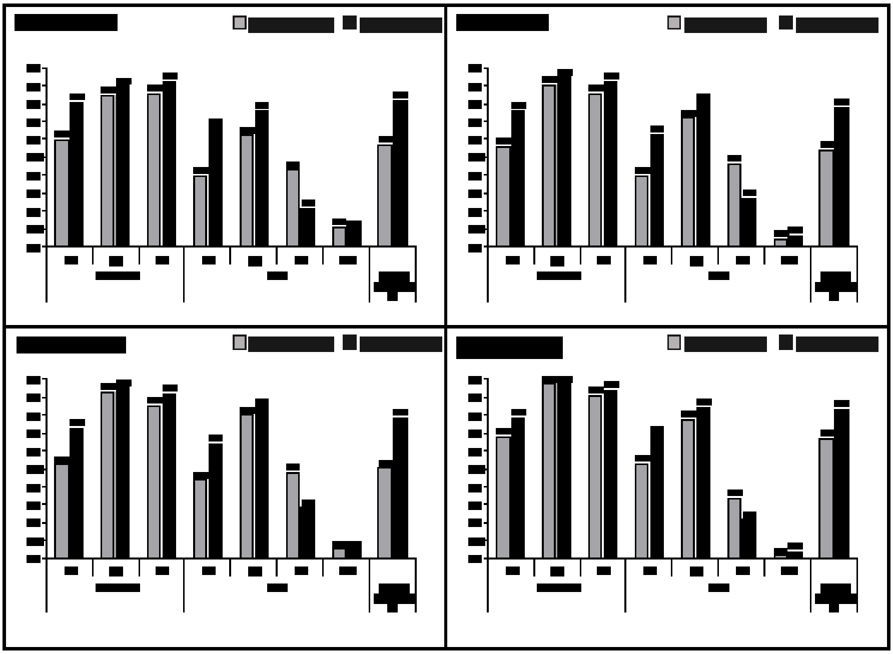


Figure 6: Prediction of small and large genes by Genscan, an ab-initio HMM based gene prediction program (Burge and Karlin 1997); GeneID, an ab-initio prediction programs that combines signals in DNA with an HMM for identifying coding sequences (Parra et al. 2000); Twinscan, a program that predicts genes in a manner similar to Genscan, but takes advantage of conserved regions between human and mouse to improve gene predictions (Korf et al. 2001); sgpGene, a program that combines GeneID predictions with TBLASTX searches between two genomes to improve predictions (Parra et al. 2003).Evaluation parameter were computed according to Burset and Guigo (Burset and Guigo 1996). Sensitivity (Sn) is the proportion of coding nucleotides/exons that were correctly predicted as coding; Specificity (Sp) is the proportion of predicted coding nucleotides/exons that are actually coding; the correlation coefficient (CC) is a global measure of accuracy combining both Sn and Sp into one value. At the exon level missing exons (ME) are exons that were completely missed by the prediction program and wrong exons (WE) are exons that were wrongly predicted where no annotated exon existed. Protein similarity was assessed by comparing the annotated protein sequence to the predicted protein sequence using BL2SEQ (Tatusova and Madden 1999).

3.3 Small human proteins

A large proportion of annotated small human proteins are known to be secreted, while the remainder was equally distributed among membrane, nuclear, mitochondria or cytoplasmic proteins (Figure 7A). Annotations suggested that human small genes were primarily involved in signal transduction and immune responses, including inflammation (Figure 7B). Furthermore, analysis of keywords from Swissprot annotations showed several protein families containing multiple small proteins (Figure 7C). The top ranking protein family (12 proteins) was the CC chemokines family. CC chemokines are chemo-attractant cytokines with two adjacent cysteines near the amino acid terminus. As small positively charged proteins, they promote the migration of monocytes in response to bacterial products, viruses and agents that cause physical damage (Cyster 1999). CC chemokines bind to CC chemokines receptors, which are G-protein coupled receptors. One other small protein family was the G-protein γ subunit family, with 9 protein members in the gene set. G-proteins coupled receptors are the largest family of cell-surface receptors and are found in all eukaryotes. G-proteins are composed of three protein subunits: α, β and γ. Under resting conditions the three subunits form a complex in which the α subunit is bound to GDP and the γ subunit is bound to the β subunit (Lambright et al. 1996). Following receptor activation the α subunit and a βγ complex (Gilman 1987) are released into the cytosol. The βγ subunits plays a prominent role in both effector regulation and receptor recognition (Clapham and Neer 1993; Haga and Haga 1992; Phillips and Cerione 1992; Pitcher et al. 1992; Tang and Gilman 1991; Wickman et al. 1994).

Another family of small proteins was the S100 family. S100 proteins are small, acidic proteins of 10-12 kDa which contain two distinct EF-hands (Ca^{2+} binding domain). These proteins act intracellularly as Ca^{2+}-signaling or Ca^{2+}-buffering proteins. Some of the S100 proteins are secreted and act in a cytokine-like manner. S100 proteins are multifunctional and are involved in the regulation of diverse cellular processes such as contraction, motility, cell growth, differentiation, cell cycle progression, transcription and secretion (for review on the S100 family see Marenholz et al. 2004). Another important small-gene family is the defensins, again with 9 members. Defensins are antimicrobial polypeptides of 12-50 amino acids and are highly abundant in cells and tissues that are involved in host defense against microbial infections. The highest concentration of defensins is found

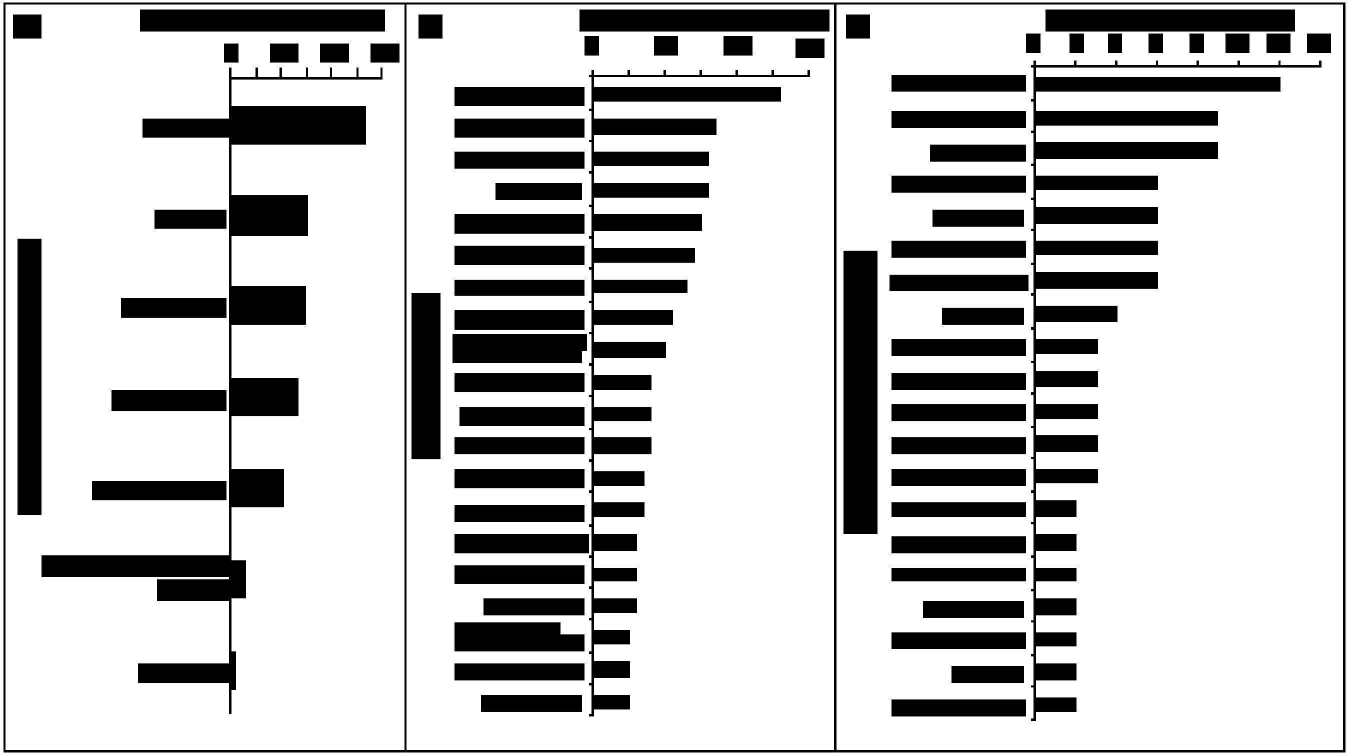


Figure 7: Annotations of small human proteins based on word count. Distribution of subcellular location of small human proteins as annotated in Swissprot (A); distribution of top 20 cell processes in which small proteins are involved as annotated in RefSeq using GO terms (B); classification of small genes into the top 20 families as annotated in Swissprot (C).

in granules of leukocytes which target infectious microorganisms (for a review see Ganz 2003 and Schneider et al. 2005).

Most of the known small genes in the human genome were discovered by studying the protein, rather than by identifying the gene. G-proteins have been studied for decades and even recently discovered proteins such as the CC-chemokines have been identified by cytokine assays. Furthermore, a relatively large proportion of the small proteins sequences (13%) was resolved by direct protein sequencing.

3.4 Expression levels of small genes

ESTs are produced by a random sampling of mRNA from tissue. Despite the bias in the type of tissue used (some tissues, such as brain, are studied more than others), EST databases such as NCBI dbEST are currently large enough to give an approximation of the level of expression of a gene. This is achieved by examining the number of ESTs from unnormalized cDNA libraries that are mapped to an annotated gene structure. Muscle glyceraldehyde-3-phosphate dehydrogenase (GAPDH), for instance, a highly abundant housekeeping gene, is represented by nearly 20,000 ESTs, while a regulatory DNA binding protein, such as *RFX5*, is represented by about 100 ESTs.

One of the difficulties in identification of small genes in bacteria was their relatively low abundance (Rudd et al. 1998; Wasinger and Humphery-Smith 1998). However, by matching ESTs to known small human genes it appears that the majority of small genes is expressed in similar amounts as all other genes (Figure 8). This observation is in accordance with the numerous cellular roles and processes in which small human proteins were shown to be involved. For instance, G-protein γ subunits must be produced in large quantities on cell surfaces. Defensins are produced in large amounts since they are stored in vesicles and excreted in bulk at a time of need; and CC cytokines are also highly abundant especially upon activation of the immune system.

4. SUMMARY AND DISCUSSION

Small proteins that are the direct product of a small genes have been primarily studied in bacteria and yeast. However, such proteins exist in humans, possibly in much higher numbers than are currently known.

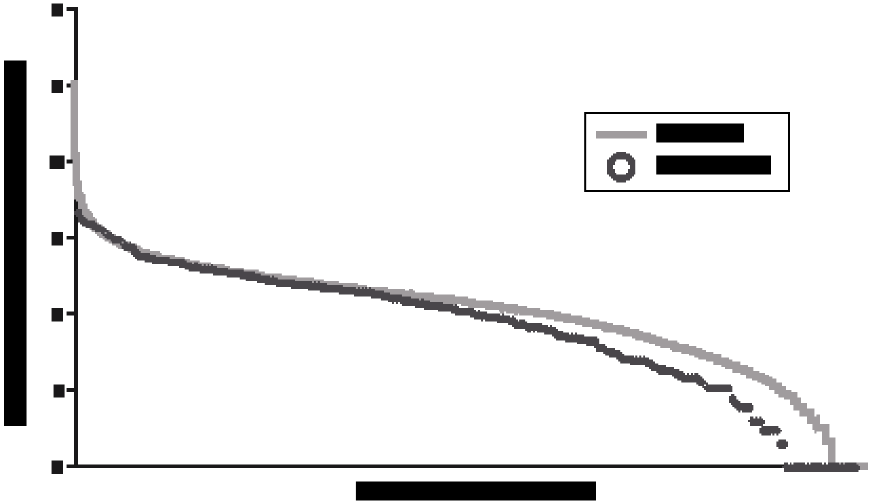


Figure 8: EST distribution of all genes (8,235 genes) and small genes (228 genes) described in Figure 1 on a logarithmic scale. The X axis represents the number of genes in each group in percentage. Human ESTs from dbEST (www.ncbi.nlm.nih.gov/dbEST) were mapped to the human genome (version HG17) using BLAT (Kent 2002). Regions identified by BLAT were further processed for enhanced accuracy using sim4 (Florea et al. 1998). ESTs with at least 50% coverage in which there was at least 90% identity to the genomic sequence were used. ESTs include both spliced and non-spliced ESTs matching in coding sequence regions.

For the past few years and especially since the announcement of the human genome sequence, annotation of the human genome and proteome has been increasingly important. The field has seen significant improvement in computational sequence analysis and in the development of novel biological methods. However, the task of annotation is far from complete. Tasks that present a great challenge and extend over a long period of time are often completed gradually, solving easy and more manageable issues first and more difficult issues later as technology improves. It is therefore natural that genes that stand out would be the first to be characterized. These include, highly expressed genes, genes that are easily identified computationally, because they include strong DNA signals and a strong codon bias, and genes that encode proteins of important function. Small genes follow the same logic. Many of the well-annotated small human genes are related to the immune system or to G-protein coupled receptors. Both have been extensively studied by the scientific community and pharmaceutical industry. Furthermore, a relatively high proportion of the known small genes were detected from the protein end rather than the genome end. This is surprising given the wealth of genomic data and the difficulties described above for identifying low molecular weight proteins. These observation suggest that the currently known small genes are those that stand out and probably many more remain to be found.

The difficulty of finding and annotating small genes extends from prediction to biological identification at both DNA and protein levels. The majority of small human genes have multiple exons, shorter in length than exons of large genes. In small genes with 3 or more exons most of the coding sequence is often contained within one internal exon and, in most cases, one exon encodes the majority of the small protein suggesting that at least one exon should be easier to predict. However, such exons can easily be misclassified as a pseudoexon. Small coding exons make small genes much harder to identify correctly using computational sequence analysis. Selected existing gene prediction programs have shown reduced sensitivity in their ability to predict small genes. Small genes are easier to identify in organisms with contiguous genes, such as bacteria and yeast. Gene prediction programs, such as sgpGene, which use alignments to known proteins, were shown to perform best at small gene detection due to their reliance on evolutionary conservation.

Human genome research has reached the stage where adequate tools and data make it possible to look deeper into the genomic sequence and identify genes and gene classes that have, thus far, escaped detection. Recent improvements include more fully sequenced genomes, allowing for better identification of conserved sequences; larger full length cDNA libraries of increased quality (Bono et al. 2002), allowing identification of more complex gene structures and alternative splicing; and large-scale microarray experiments that identify transcribed regions along an entire chromosome (Cheng et al. 2005). The availability of such data paves the way to identify and annotate more small genes, however, those genes will always be harder to find statistically and any method based on random selection of mRNA and/or proteins from tissue is likely to be inadequate. Small genes will need to be specifically targeted. One such example was given by Reymond et al. 2002, who combined *in silico* and biological evidence to find gene candidates, followed by 3′ and 5′ RACE to find and annotate genes(Reymond et al. 2002). Other projects have used targeted gene finding to identify novel, previously unannotated human genes (ENCODE project consortium 2004; Dike et al. 2004; Guigo et al. 2003). Small genes should first be computationally identified by combining several sets of evidence: gene prediction data from various sources (the combination of several programs has been shown more effective than any individual program (Pavlovic et al. 2002; Yada et al. 2003)); conserved sequences between the human genome and other vertebrate genomes; alignments of known proteins to the human genome; and experimental data such as ESTs, transcriptomics and proteomics. The combination of such data is likely to increase the probability of identifying at least one exon per small gene and must be carefully analyzed to exclude pseudogenes. Potential small gene candidates should be further validated using 3′ and 5′ RACE with specific primers. The full mRNA sequence can be derived experimentally in two steps: a 3′ RACE from a poly (A) binding primer to a specific primer in the identified exon; and a 5′ RACE from a cap binding primer to a specific primer in the identified exon.

Human genome annotation is at a stage where random biologically collected data with respect to gene finding contributes limited new information. Small genes that have been generally ignored as a class can now be targeted computationally and experimentally so as to advance human genome annotation one step further.

The next phase of human gene annotation will revert increasingly to a reliance on experimental correction/validation, particularly of start codon and exon structure. For smaller genes, proteomics may prove the more efficient means to clarify gene structure. As a more ordered approach to the Human Proteome Project is heralded (Humphery-Smith 2004), the Protein Atlas (Nilsson et al. 2005; Uhlen et al. 2005) will collate experimentally-observed parameters to validated gene structure. Elsewhere, the utility of datasets outlining previously-observed peptides during high-throughput proteomics should lend credence to exon structure as a means to better annotate human genomic sequence.

ACKNOWLEDGMENTS

Work in Zhang lab was supported by grants from NIH. We thank Rob Laveaux for donating his “Scripting Tools” software, a plug-in for 4th dimension database.

REFERENCES

Ashurst, J.L., C.K. Chen, J.G. Gilbert, K. Jekosch, S. Keenan, P. Meidl, S.M. Searle, J. Stalker, R. Storey, S. Trevanion et al. 2005. The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res* **33:** D459-465.

Bairoch, A., R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane et al. 2005. The Universal Protein Resource (UniProt). *Nucleic Acids Res* **33 :** D154-159.

Balakirev, E.S. and F.J. Ayala. 2003. Pseudogenes: are they "junk" or functional DNA? *Annu Rev Genet* **37:** 123-151.

Barry, C., G. Fichant, A. Kalogeropoulos, and Y. Quentin. 1996. A computer filtering method to drive out tiny genes from the yeast genome. *Yeast* **12:** 1163-1178.

Basrai, M.A., P. Hieter, and J.D. Boeke. 1997. Small open reading frames: beautiful needles in the haystack. *Genome Res* **7:** 768-771.

Bell, G.I., R.L. Pictet, W.J. Rutter, B. Cordell, E. Tischer, and H.M. Goodman. 1980. Sequence of the human insulin gene. *Nature* **284:** 26-32.

Bono, H., T. Kasukawa, M. Furuno, Y. Hayashizaki, and Y. Okazaki. 2002. FANTOM DB: database of Functional Annotation of RIKEN Mouse cDNA Clones. *Nucleic Acids Res* **30:** 116-118.

Burge, C. and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268:** 78-94.

Burset, M. and R. Guigo. 1996. Evaluation of gene structure prediction programs. *Genomics* **34:** 353-367.

Cebrat, S., P. Mackiewicz, and M.R. Dudek. 1998. The role of the genetic code in generating new coding sequences inside existing genes. *Biosystems* **45:** 165-176.

Cheng, J., P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308:** 1149-1154.

Clapham, D.E. and E.J. Neer. 1993. New roles for G-protein βγ-dimers in transmembrane signalling. *Nature* **365:** 403-406.

Coin, L. and R. Durbin. 2004. Improved techniques for the identification of pseudogenes. *Bioinformatics* **20 Suppl 1:** I94-I100.

Collins, J.E., M.E. Goward, C.G. Cole, L.J. Smink, E.J. Huckle, S. Knowles, J.M. Bye, D.M. Beare, and I. Dunham. 2003. Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Res* **13:** 27-36.

Cyster, J.G. 1999. Chemokines and cell migration in secondary lymphoid organs. *Science* **286:** 2098-2102.

Dike, S., V.S. Balija, L.U. Nascimento, Z. Xuan, J. Ou, T. Zutavern, L.E. Palmer, G. Hannon, M.Q. Zhang, and W.R. McCombie. 2004. The mouse genome: experimental examination of gene predictions and transcriptional start sites. *Genome Res* **14:** 2424-2429.

Durbin, R., S.R. Eddy, A. Krogh, and G. Mitchison. 1998. Biological sequence analysis, pp. 46-79. Cambridge University Press, Cambridge.

Eddy, S.R. 2001. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet* **2:** 919-929.

ENCODE project consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636-640.

Fickett, J.W. 1994. Inferring genes from open reading frames. *Comput Chem* **18:** 203-205.

Fickett, J.W. 1995. ORFs and genes: how strong a connection? *J Comput Biol* **2:** 117-123.

Florea, L., G. Hartzell, Z. Zhang, G.M. Rubin, and W. Miller. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* **8:** 967-974.

Ganz, T. 2003. Defensins: antimicrobial peptides of innate immunity. *Nat Rev Immunol* **3:** 710-720.

Gao, F. and C.T. Zhang. 2004. Comparison of various algorithms for recognizing short coding sequences of human genes. *Bioinformatics* **20:** 673-681.

Gibbs, R.A. G.M. Weinstock M.L. Metzker D.M. Muzny E.J. Sodergren S. Scherer G. Scott D. Steffen K.C. Worley P.E. Burch et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493-521.

Gilman, A.G. 1987. G proteins: transducers of receptor-generated signals. *Annu Rev Biochem* **56:** 615-649.

Guigo, R., P. Agarwal, J.F. Abril, M. Burset, and J.W. Fickett. 2000. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res* **10:** 1631-1642.

Guigo, R., E.T. Dermitzakis, P. Agarwal, C.P. Ponting, G. Parra, A. Reymond, J.F. Abril, E. Keibler, R. Lyle, C. Ucla et al. 2003. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc Natl Acad Sci U S A* **100:** 1140-1145.

Gygi, S.P., G.L. Corthals, Y. Zhang, Y. Rochon, and R. Aebersold. 2000. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci U S A* **97:** 9390-9395.

Haga, K. and T. Haga. 1992. Activation by G protein β γ subunits of agonist- or light-dependent phosphorylation of muscarinic acetylcholine receptors and rhodopsin. *J Biol Chem* **267:** 2222-2227.

HGSC. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931-945.

Humphery-Smith, I. 2004. A human proteome project with a beginning and an end. *Proteomics* **4:** 2519-2521.

Jansen, R., H.J. Bussemaker, and M. Gerstein. 2003. Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic Acids Res* **31:** 2242-2251.

Karolchik, D., R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res* **31:** 51-54.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664.

Koonin, E.V., R. Tatusov, and K.E. Rudd. 1997. In *Escherichia coli and Salmonella, Cellular and Molecular Biology* (eds. F.C. Neidhardt R. Curtiss J.L. Ingraham E.C.C. Lin K.B. Low B. Magasanik W.S. Reznikoff M. Riley M. Schaechter, and H.E. Umbarger), pp. 2203-2217. ASM Press, Wasington, DC.

Korf, I., P. Flicek, D. Duan, and M.R. Brent. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17 Suppl 1:** S140-148.

Krogh, A. 1997. Two methods for improving performance of an HMM and their application for gene finding. *Proc Int Conf Intell Syst Mol Biol* **5:** 179-186.

Krogh, A. 2000. Using database matches with for HMMGene for automated gene detection in Drosophila. *Genome Res* **10:** 523-528.

Kulp, D., D. Haussler, M.G. Reese, and F.H. Eeckman. 1996. A generalized hidden Markov model for the recognition of human genes in DNA. *Proc Int Conf Intell Syst Mol Biol* **4:** 134-142.

Lambright, D.G., J. Sondek, A. Bohm, N.P. Skiba, H.E. Hamm, and P.B. Sigler. 1996. The 2.0 A crystal structure of a heterotrimeric G protein. *Nature* **379:** 311-319.

Lazazzera, B.A., J.M. Solomon, and A.D. Grossman. 1997. An exported peptide functions intracellularly to contribute to cell density signaling in B. subtilis. *Cell* **89:** 917-925.

Liu, Y., P.M. Harrison, V. Kunin, and M. Gerstein. 2004. Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome Biol* **5:** R64.

Mackiewicz, P., M. Kowalczuk, A. Gierlik, M.R. Dudek, and S. Cebrat. 1999. Origin and properties of non-coding ORFs in the yeast genome. *Nucleic Acids Res* **27:** 3503-3509.

Mackiewicz, P., M. Kowalczuk, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, A. Laszkiewicz, M.R. Dudek, and S. Cebrat. 2002. How many protein-coding genes are there in the Saccharomyces cerevisiae genome? *Yeast* **19:** 619-629.

Mann, M., R.C. Hendrickson, and A. Pandey. 2001. Analysis of proteins and proteomes by mass spectrometry. *Annu Rev Biochem* **70:** 437-473.

Marenholz, I., C.W. Heizmann, and G. Fritz. 2004. S100 proteins in mouse and man: from evolution to function and pathology (including an update of the nomenclature). *Biochem Biophys Res Commun* **322:** 1111-1122.

Martin, S.E., J. Shabanowitz, D.F. Hunt, and J.A. Marto. 2000. Subfemtomole MS and MS/MS peptide sequence analysis using nano-HPLC micro-ESI fourier transform ion cyclotron resonance mass spectrometry. *Anal Chem* **72:** 4266-4274.

Mewes, H.W., C. Amid, R. Arnold, D. Frishman, U. Guldener, G. Mannhaupt, M. Munsterkotter, P. Pagel, N. Strack, V. Stumpflen et al. 2004. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res* **32:** D41-44.

Misra, S., M.A. Crosby, C.J. Mungall, B.B. Matthews, K.S. Campbell, P. Hradecky, Y. Huang, J.S. Kaminker, G.H. Millburn, S.E. Prochnik et al. 2002. Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. *Genome Biol* **3:** RESEARCH0083.

Nilsson, P., L. Paavilainen, K. Larsson, J. Odling, M. Sundberg, A.C. Andersson, C. Kampf, A. Persson, C. Al-Khalili Szigyarto, J. Ottosson et al. 2005. Towards a human proteome atlas: high-throughput generation of mono-specific antibodies for tissue profiling. *Proteomics* **5:** 4327-4337.

Oliver, S.G., Q.J. van der Aart, M.L. Agostoni-Carbone, M. Aigle, L. Alberghina, D. Alexandraki, G. Antoine, R. Anwar, J.P. Ballesta, P. Benit et al. 1992. The complete DNA sequence of yeast chromosome III. *Nature* **357:** 38-46.

Pandey, A. and F. Lewitter. 1999. Nucleotide sequence databases: a gold mine for biologists. *Trends Biochem Sci* **24:** 276-280.

Parra, G., P. Agarwal, J.F. Abril, T. Wiehe, J.W. Fickett, and R. Guigo. 2003. Comparative gene prediction in human and mouse. *Genome Res* **13:** 108-117.

Parra, G., E. Blanco, and R. Guigo. 2000. GeneID in Drosophila. *Genome Res* **10:** 511-515.

Pavlovic, V., A. Garg, and S. Kasif. 2002. A Bayesian framework for combining gene predictions. *Bioinformatics* **18:** 19-27.

Phillips, W.J. and R.A. Cerione. 1992. Rhodopsin/transducin interactions. I. Characterization of the binding of the transducin-βγ subunit complex to rhodopsin using fluorescence spectroscopy. *J Biol Chem* **267:** 17032-17039.

Pitcher, J.A., J. Inglese, J.B. Higgins, J.L. Arriza, P.J. Casey, C. Kim, J.L. Benovic, M.M. Kwatra, M.G. Caron, and R.J. Lefkowitz. 1992. Role of βγ subunits of G proteins in targeting the β-adrenergic receptor kinase to membrane-bound receptors. *Science* **257:** 1264-1267.

Rabilloud, T. 2000. *Proteome Research: Two-Dimensional Gel Electrophoresis and Identification Methods*. Springer, Berlin.

Reymond, A., A.A. Camargo, S. Deutsch, B.J. Stevenson, R.B. Parmigiani, C. Ucla, F. Bettoni, C. Rossier, R. Lyle, M. Guipponi et al. 2002. Nineteen additional unpredicted transcripts from human chromosome 21. *Genomics* **79:** 824-832.

Rogic, S., A.K. Mackworth, and F.B. Ouellette. 2001. Evaluation of gene-finding programs on mammalian sequences. *Genome Res* **11:** 817-832.

Rudd, K.E., I. Humphery-Smith, V.C. Wasinger, and A. Bairoch. 1998. Low molecular weight proteins: a challenge for post-genomic research. *Electrophoresis* **19:** 536-544.

Salzberg, S., A.L. Delcher, K.H. Fasman, and J. Henderson. 1998. A decision tree system for finding genes in DNA. *J Comput Biol* **5:** 667-680.

Sambrook, J. and D. Russell. 2001. Molecular cloning: a laboratory manual, pp. 11.57. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Schneider, J.J., A. Unholzer, M. Schaller, M. Schafer-Korting, and H.C. Korting. 2005. Human defensins. *J Mol Med*.

Schuler, G.D. 1997. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J Mol Med* **75:** 694-698.

Schuler, G.D. M.S. Boguski E.A. Stewart L.D. Stein G. Gyapay K. Rice R.E. White P. Rodriguez-Tome A. Aggarwal E. Bajorek et al. 1996. A gene map of the human genome. *Science* **274:** 540-546.

Sharp, P.M. and E. Cowe. 1991. Synonymous codon usage in Saccharomyces cerevisiae. *Yeast* **7:** 657-678.

Siepel, A. and D. Haussler. 2004. Computational identification of evolutionarily conserved exons. In *Proceedings of the eighth annual international conference on Computational molecular biology*, pp. 177-186. ACM Press, San Diego, California, USA.

Solovyev, V.V., A.A. Salamov, and C.B. Lawrence. 1995. Identification of human gene structure using linear discriminant functions and dynamic programming. *Proc Int Conf Intell Syst Mol Biol* **3:** 367-375.

Solovyev, V.V. 2001. Statistical approaches in Eukaryotic gene prediction. In *Handbook of Statistical Genetics* (eds. D.J. Balding C. Cannings, and B. M.), pp. 83-128. John Wiley & Sons.

Stein, L.D., C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J.E. Stajich, T.W. Harris, A. Arva et al. 2002. The generic genome browser: a building block for a model organism system database. *Genome Res* **12:** 1599-1610.

Stormo, G.D. 2000. Gene-finding approaches for eukaryotes. *Genome Res* **10:** 394-397.

Takahashi, H., Y. Teranishi, S. Nakanishi, and S. Numa. 1981. Isolation and structural organization of the human corticotropin-β-lipotropin precursor gene. *FEBS Lett* **135:** 97-102.

Tang, W.J. and A.G. Gilman. 1991. Type-specific regulation of adenylyl cyclase by G protein βγ subunits. *Science* **254:** 1500-1503.

Tatusova, T.A. and T.L. Madden. 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* **174:** 247-250.

Torrents, D., M. Suyama, E. Zdobnov, and P. Bork. 2003. A genome-wide survey of human pseudogenes. *Genome Res* **13:** 2559-2567.

Uhlen, M., E. Bjorling, C. Agaton, C.A. Szigyarto, B. Amini, E. Andersen, A.C. Andersson, P. Angelidou, A. Asplund, C. Asplund et al. 2005. A Human Protein Atlas for Normal and Cancer Tissues Based on Antibody Proteomics. *Mol Cell Proteomics* **4:** 1920-1932.

Wasinger, V.C., S.J. Cordwell, A. Cerpa-Poljak, J.X. Yan, A.A. Gooley, M.R. Wilkins, M.W. Duncan, R. Harris, K.L. Williams, and I. Humphery-Smith. 1995. Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. *Electrophoresis* **16:** 1090-1094.

Wasinger, V.C. and I. Humphery-Smith. 1998. Small genes/gene-products in *Escherichia coli* K-12. *FEMS Microbiol Lett* **169:** 375-382.

Waterston, R.H. K. Lindblad-Toh E. Birney J. Rogers J.F. Abril P. Agarwal R. Agarwala R. Ainscough M. Alexandersson P. An et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.

Wickman, K.D., J.A. Iniguez-Lluhl, P.A. Davenport, R. Taussig, G.B. Krapivinsky, M.E. Linder, A.G. Gilman, and D.E. Clapham. 1994. Recombinant G-protein βγ-subunits activate the muscarinic-gated atrial potassium channel. *Nature* **368:** 255-257.

Yada, T. and M. Hirosawa. 1996. Detection of short protein coding regions within the cyanobacterium genome: application of the hidden Markov model. *DNA Res* **3:** 355-361.

Yada, T., T. Takagi, Y. Totoki, Y. Sakaki, and Y. Takaeda. 2003. DIGIT: a novel gene finding program by combining gene-finders. *Pac Symp Biocomput***:** 375-387.

Yates, J.R., 3rd, A.L. McCormack, D. Schieltz, E. Carmack, and A. Link. 1997. Direct analysis of protein mixtures by tandem mass spectrometry. *J Protein Chem* **16:** 495-497.

Zhang, C.T. and R. Zhang. 2002. Evaluation of gene-finding algorithms by a content-balancing accuracy index. *J Biomol Struct Dyn* **19:** 1045-1052.

Zhang, M.Q. 1997. Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc Natl Acad Sci U S A* **94:** 565-568.

Zhang, M.Q. 2002. Computational prediction of eukaryotic protein-coding genes. *Nat Rev Genet* **3:** 698-709.

Zhang, Z. and M. Gerstein. 2004. Large-scale analysis of pseudogenes in the human genome. *Curr Opin Genet Dev* **14:** 328-335.

Zhang, Z., P.M. Harrison, Y. Liu, and M. Gerstein. 2003. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res* **13:** 2541-2558.

3

Prediction of small protein-coding genes in the human genome

Yair Benita[1], Zhenyu Xuan[2], Michael J Wise[3], Michael Q Zhang[2], Ian Humphery-Smith[4] and Ronald S Oosting[1]

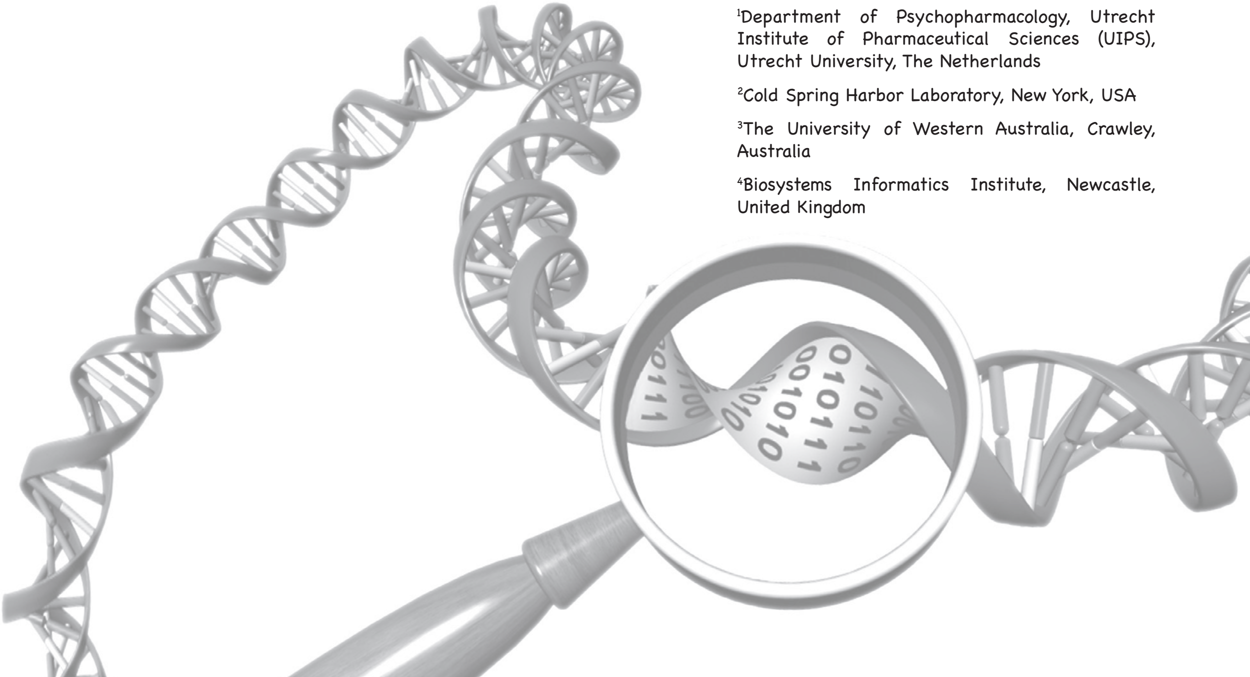
Manuscript in preparation

[1]Department of Psychopharmacology, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, The Netherlands

[2]Cold Spring Harbor Laboratory, New York, USA

[3]The University of Western Australia, Crawley, Australia

[4]Biosystems Informatics Institute, Newcastle, United Kingdom



ABSTRACT

Small genes that encode a protein smaller than 100 amino acids fall in the gray area of both genomics and proteomics. Genome annotation is systematically biased against annotating small genes due to their lower probability of being real compared to large genes; and proteomics is in need for more sensitive and accurate methods to separate and identify the low molecular weight proteome.

Gene prediction programs, evolutionary conserved elements and homology to ESTs, known proteins and protein motifs were combined into an algorithm biased toward finding small genes. These identified gene candidates were filtered to exclude pseudogenes, viral and transposable elements, resulting in a prediction set of 1,665 multi-exon and 7,709 single exon non-annotated small genes.

INTRODUCTION

With the sequence of the human genome now largely determined attention has shifted towards genome annotation, a process that identifies sequence features on genomic DNA. This process includes the identification of protein coding-genes, non-coding genes, binding sites for transcription factors and other regulatory elements and eventually elucidating the biological function of these elements.

The current human genome annotation, with respect to gene finding, is a combination of three approaches: (i) gene prediction programs; (ii) mapping known expressed sequence tags (ESTs) and proven genes (protein-coding and RNA) onto genomic sequences; and (iii) identification of conserved sequences from other vertebrate genomes that are likely to be genes. Using this annotation strategy the human genome sequencing consortium estimated the total number of protein-coding genes in the human genome to be in the range of 20,000-25,000 genes (HGSC 2004). This range of 5000 human genes is attributed mostly to: (i) genes with very small ORFs of less than 100 amino acids; (ii) single exon genes; and (iii) genes that evolved rapidly. In this study we focused on the identification of small protein-coding genes in the human genome. Here, the term "ORF" is employed to refer to a continuous sequence in DNA or RNA that can potentially encode a protein or part of a protein. The term "gene" refers to a contiguous or non-contiguous DNA coding sequence that actually encodes a protein and the term "small" refers to genes and proteins comprised of less than 100 codons or amino acids, respectively. We also employ the terms spliced EST and spliced protein to refer to an EST or protein that were mapped to genomic DNA sequence and matched in two or more non-contiguous fragments.

Small genes have been primarily studied in bacteria and yeast (Barry et al. 1996; Yada and Hirosawa 1996; Wasinger and Humphery-Smith 1998). The identification of these genes, due to their size, was shown to be challenging both on the computational side and on the biological side (Basrai et al. 1997; Rudd et al. 1998; Wasinger and Humphery-Smith 1998). In chapter 2, a detailed overview of 228 currently annotated small human genes was presented. The main findings of this overview can be summarized as follows: (i) small genes had on average 2-3 coding exons, up to 7; (ii) coding exons of small genes were on average half the size that of coding exons of large genes; (iii) in multi-exon small genes most of the coding sequence was contained within one exon; (iv) gene prediction programs showed reduced sensitivity in the prediction of small genes; (v) expression levels of small annotated genes appeared to be similar to expression levels of large genes.

Current genome annotation strategies rely heavily on expressed sequence tags (ESTs) and cDNA libraries. This was demonstrated by the impact of increasing numbers of ESTs and cDNAs on the annotation of the *Drosophila melanogaster* genome (Misra et al. 2002). In revision 3 of the annotation as many as 85% of the annotated gene models of revision 2 were altered and 273 novel genes were added. The contribution of ESTs and cDNAs to the human genome annotation has also been significant. These sequences were obtained using random sequencing of mRNA from tissue. Currently, dbEST (Boguski et al. 1993) contains over 7 million human sequences and despite the constant growth very little new information is being added with respect to gene finding. In a recent paper announcing the completion of the euchromatic sequence of the human genome the authors mentioned that "despite intense automated and manual analysis using cDNAs, ESTs and cross-species homology, only 2,188 gene predictions have been added to the known set" (HGSC 2004). Therefore, it is reasonable to assume that small genes that were not discovered so far are not likely to be discovered by random sequencing of mRNA. This may partly be due to the way the cDNA libraries were created, i.e. only mRNAs larger than a certain size (e.g. 1,000 bp) were reverse transcribed.

Despite the high quality of the human genome sequence, advances in gene prediction algorithms and the accumulation of genomic driven data, many of the annotated small genes were discovered from the protein end (chapter 2). Examples include CC chemokines, a family of chemo-attractant cytokines that was discovered by immunoassays (Cyster 1999); G-protein coupled receptors γ subunit family; defensins, a family of antimicrobial polypeptides (Ganz 2003); and S100, a family of small calcium binding acidic proteins (Marenholz et al. 2004). The current human genome annotation strategy is not suitable for the identification of such small genes as they often lie below the statistical significance threshold (see chapter 2). However, current annotation strategy can be biased to identify exons of small genes that stand-out and these sequences can be used to generate specific primers and extract the full mRNA sequence from tissue. In this study we outline the computational tools and data employed to identify potential new small genes. We trained and estimated the reliability of the algorithms on a set of 228 annotated small genes and present the results of our search for new small genes across the entire human genome. Finally, we suggest a PCR based method to confirm the existence of the predicted small genes and to discover the entire mRNA sequence. The small genes database, sequences and genome browser are available at http://ruali.cshl.edu/smallgenes.

METHODS

Reference set of annotated protein coding genes

A well annotated set of human genes was compiled for training and testing purposed from SwissProt (Bairoch et al. 2005) and the Consensus Coding Sequence database (CCDS) (http://www.ncbi.nlm.nih.gov/CCDS). Human sequences from CCDS that matched a protein sequence from SwissProt with an identity and sequence coverage above 95% were selected. Other gene annotations that were used as a reference but not included the training or testing datasets included sequences from The Vertebrate Genome Annotation database (Vega) (Ashurst et al. 2005), sequences from Reference Sequence (RefSeq) (Pruitt et al. 2005) and the known gene track from the UCSC genome browser (Kent et al. 2002).

Data used in the identification of protein coding genes

The genome revisions that were used in the analysis for human, mouse and rat were HG17, MM6 and RN3, respectively. Genome data and *axt* pairwise-alignments produced using Blastz were downloaded from the UCSC genome web server (Karolchik et al. 2003). Gene prediction programs used included: (i) Genscan, an ab-initio HMM based program (Burge and Karlin 1997); (ii) Twinscan, a program that predicts genes in a manner similar to Genscan, but takes advantage of conserved regions between human and mouse to improve gene predictions (Korf et al. 2001); (iii) GeneID, an ab-initio prediction programs that combines signals in DNA with an HMM for identifying coding sequences (Parra et al. 2000); (iv) sgpGene, a program that combines GeneID predictions with TBLASTX searches between two genomes to improve predictions (Parra et al. 2003); (v) Acembly, a genome annotation program based on EST data (Thierry-Mieg and Thierry-Mieg); and (vi) Ensembl, an automated annotation system based on homology to known proteins, cDNAs and ESTs (Hubbard et al. 2002; Curwen et al. 2004).

Homology searches

Candidate exons were compared to the NCBI non-redundant protein database excluding all sequences with the keywords “hypothetical” and “predicted” using BlastX (Altschul et al. 1990) with an e-value of 0.01. Overlapping protein hits on the same frame were clustered and the full protein sequence of the top 5 protein matches of each cluster were realigned to the relevant chromosome using BLAT (Kent 2002). Only proteins with non-contiguous matches were selected from this analysis and are referred to as spliced proteins. The presence of these proteins was used both in the logistic regression model and to support a connection between two predicted exons. Comparison to known motifs was performed using RPS-BLAST (Altschul et al. 1997) and the Conserved Domain Database (CDD) (Marchler-Bauer and Bryant 2004) version 2.0.5 with an e-value of 0.1. CDD included Pfam (Bateman et al. 2004), SMART (Letunic et al. 2004), COGs (Tatusov et al. 2003), KOGs (Tatusov et al. 2003) and NCBI curated motifs. Additional comparison to protein motifs was exclusively performed for chromosome 22 revision HG16 using InterProScan version 3.3 to query the InterPro database version 7.0 (Mulder et al. 2003). Each CCP was subject to a self-homology test by comparing the sequence to the entire human genome using BLAT (Kent 2002) with default parameters. Human ESTs from dbEST (Boguski et al. 1993) were aligned to the human genome using BLAT (Kent 2002) and the regions identified by BLAT were further processed for enhanced accuracy using sim4 (Florea et al. 1998). ESTs with at least 50% coverage in which there was at least 90% identity to the genomic sequence were used. ESTs were separated to spliced and non-spliced and each group was clustered to form the largest possible gene on each strand. The edges of contiguous fragments of spliced ESTs were determined using the chromosomal position of the majority of the ESTs in the cluster that were spliced at that edge.

Transcriptome data

Affymetrix transcription data from tiling microarrays was used to estimate chromosomal transcription (Kapranov et al. 2002). The data used was downloaded from the UCSC web server for genome version HG16 and tested exclusively for chromosome 22. Probes were scored on a scale of 0-1000 and only probes with scores equal or greater than 100 were used, a score much higher than the original score determined to be a true positive by the authors (Kapranov et al. 2002). Chromosome 22 HG16 was further used to estimate the contribution of transcription data from Rinn et al. (Rinn et al. 2003) to gene finding. Only sequences that were confirmed by PCR were used.

Homology to other species

Sequences conserved in other species were added by using BLASTz alignments of human-dog (canFam2), human-chicken (galGal2), human-fugu (Fr1), human-tetraodon (tetNig1) and human-Zebrafish (dnaRer2) as distributed by UCSC (Karolchik et al. 2003). Furthermore the program phastCons (Siepel et al. 2005) was used to score each nucleotide for conservation.

Logistic regression model

The training and testing sets for the logistic regression were generated from all CCP elements. A combination of data sources suggesting a nucleotide was coding a protein was treated as a vector and each vector had a weight relative to the number of occurrences in each set. Large genes and small genes from the reference set described above were used as positive examples. The remaining CCP elements were used as a negative set, excluding all sequences annotated in the Vega database, UCSC known genes and RefSeq. The number of nucleotides in the negative set was extensively larger than the number of nucleotides in the positive set, since CCP elements covered 19% of the human genome and current annotations cover ~1%. Therefore, the weight of vectors in the negative set were scaled relatively so that the sum of weights of the negative set was identical to the sum of weights of the positive set. While many of the nucleotides in the negative set may be pseudogenes or protein coding genes, the vast majority is most likely not coding for proteins. Therefore, the weight of non-annotated protein-coding genes in the negative set was assumed to be insignificant. The training and testing sets included a random selection of 70% and 30% of the nucleotides, respectively, from the positive and negative sets. The same negative set was used for training and testing of both small and large genes, however, the set was shuffled randomly and split to 70% and 30% separately for each gene group.

Logistic regression was carried out using the binary logistic command of SPSS version 11 and the best model was selected based on the lowest -2log likelihood score, which is commonly used as a goodness of fit metric for logistic regression.

Sensitivity and specificity calculations

Sensitivity (the probability of correctly predicting a coding nucleotide) and specificity (the probability that a positive example is correct) and the correlation coefficient (CC) were calculated according to Burset and Guigo (Burset and Guigo 1996). The CC is a measure combining both sensitivity and specificity into one value in the range of -1 to 1.

Database

A customized database was constructed using Python programming language and ZOPE object oriented database (ZODB). The database was based on two types of structures, continuous elements, such as exons, and connection elements that hold the structure of the exons. Each database file was designed to hold data of one chromosome only and was optimized for searching quickly and efficiently overlapping elements based on their chromosomal position. The database was used to store Python objects rather than text information, allowing an efficient design of indices and methods. The database included several methods that are useful for genomic research, for example, a method for clustering overlapping elements, such as EST, and generating the most probable gene structure. The python scripts and instructions for building such a databases will be made available upon request.

RESULTS

A high quality reference gene set that contained 8,007 annotated large genes and 228 annotated small genes was compiled and used for training and testing purposes. The primary assumption in the computational detection of small genes was that at least one exon was detectable. This assumption is based on the observation that in multi-exon small genes the majority of the coding sequence is usually contained within one exon. In our reference set the average largest exon length was 131 bp in multi-exon small genes, covering 54% of the entire coding sequence. A large initial set of DNA sequences that potentially contained protein coding exons was generated and data from various sources were combined in an attempt to reduce this large set and locate exons of small genes that stand-out. The initial set of DNA sequences was generated by combining overlapping exons of gene prediction programs, mouse-human and rat-human alignments into the largest possible continuous DNA sequences. This set of elements is referred to as CCP elements (Clusters of Conserved and Predicted elements). Details of their construction and the data collected to identify coding exons are summarized in Figure 1. CCP elements that contained only human-mouse or human-rat alignments and no other EST data or gene

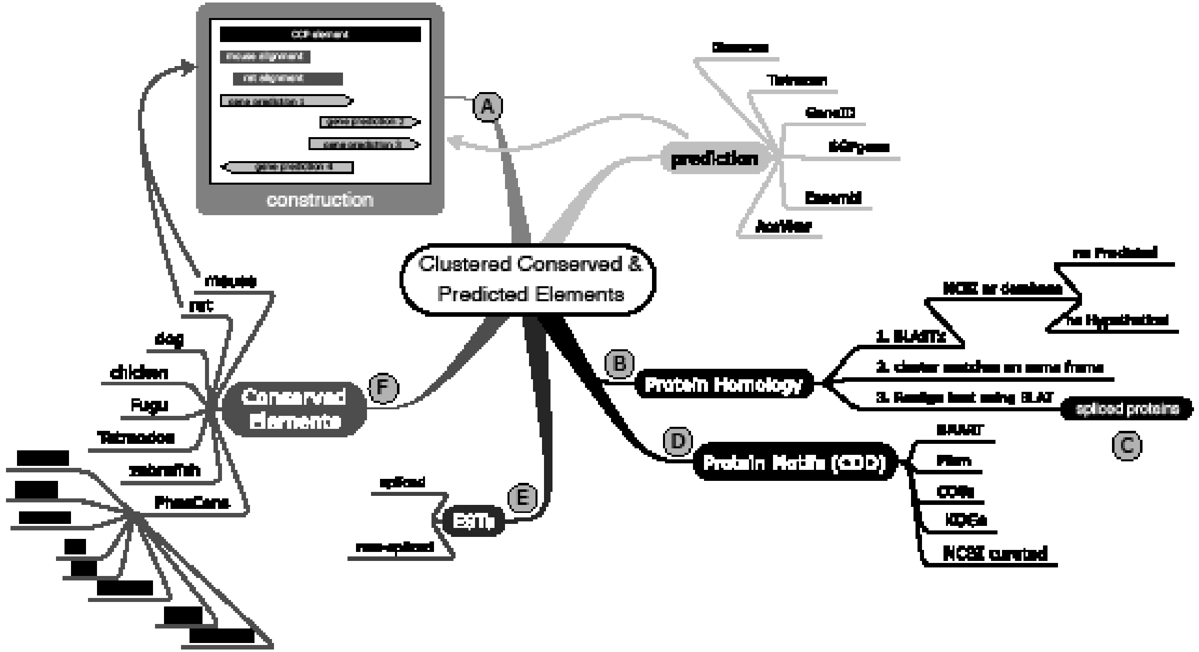


Figure 1: A mind map summarizing the construction of initial DNA sequence to be analyzed for identification of protein coding exons. Clustered conserved and predicted elements (CCP) were constructed by clustering overlapping exons of gene prediction programs, human-mouse and human-rat alignments into one contiguous sequence (A). Each of these elements was further examined for the likelihood of containing a protein coding exon. Homology to known proteins was identified using BlastX and the NCBI non-redundant (nr) protein database (B). The top five protein hits on each frame were realigned to the chromosome using BLAT in order to locate other fragment of those proteins (C). Protein motifs and patterns were identified using RPS-BLAST and the NCBI Conserved Domain Database (CDD) (D). Homology to ESTs from dbEST was assessed using BLAT and sim4 and ESTs were divided to spliced and non-spliced (E). Conserved elements, based on BlastZ genome alignments in other mammals, fish and chicken were added together with PhastCons elements (F). PhastCons is a phylogenetic hidden markov model used to identify evolutionary conserved elements in the human genome using the species outlined in the map.

prediction data suggesting the existence of a coding exon were discarded. Approximately 487,000 CCP elements were identified which covered 19% of the human genome, while the current annotations cover 0.74%, 1.2%, 1.8% and 1.4% for CCDS, RefSeq, UCSC and Vega, respectively (see methods for details on annotation databases). The identified CCP elements contained all protein coding exons from the reference set including all exons of the small genes. However, the average CCP element size was 1208 bp compared to the much smaller annotated exon size of 164 bp in the reference set. ESTs from dbEST were separated into spliced and non-spliced ESTs and considered for each strand separately. Approximately 36% and 39% of CCP elements had at least one overlapping spliced and non-spliced EST element, respectively.

Each CCP elements was compared to known proteins using BLASTX and the five best matches for each of the six translated frames were realigned to the relevant chromosome to increase confidence in coding probability and to detect possible connections between adjacent exons. In total 52% of the CCP elements had at least one match to a known protein on at least one of the six frames. Proteins that matched a CCP element and were spliced, i.e. connecting adjacent CCP elements, were considered to be more reliable for the purpose of gene finding. Slightly more than 28% of the CCP elements had homology to a protein that was spliced. Protein motifs were detected using the NCBI conserved domain database (CDD) on a six frame translation of each CCP element. Forty five percent of the CCP elements contained at least one protein motif on one of the six frames.

Human chromosome 22 (NCBI build 34) was used to estimate the contribution of Interpro analysis, transcriptome data from Affymetrix and the from Yale chromosome 22 microarray database to the identification of coding exons in CCP elements. Interpro analysis was expensive in computer power (at least 10 fold more than CDD) and generated 2028 unique non-overlapping protein motifs and patterns in 14% of the CCP elements in chromosome 22. In comparison, CDD generated 4122 unique non-overlapping protein motifs and patterns in 29% of the CCP elements. Combining both sources, about 30% of the motifs were identified both in CDD and Interpro, 60% exclusively in CDD and 10% exclusively in Interpro. Despite the possibility that the motifs identified exclusively by Interpro may contribute to the gene finding process, the additional computational power needed for the procedure could not be justified. Therefore, Interpro analysis was not performed for the rest of the human genome.

The contribution of transcriptome data to the process of gene finding was evaluated. This type of data is typically generated by using DNA microarrays containing tiled oligonucleotide probes that query the greater part of an entire chromosome. Selected Affymetrix positive probes covered 4.5% of chromosome 22 compared to 1.35% covered by Vega annotated coding exons. However, only 9.7% of the Vega annotated nucleotides were supported by transcription data from the Affymetrix probes. The vast majority of the transcription occurred in areas that were not annotated as coding exons. Transcriptome data from the Yale chromosome 22 microarray database (http://array.mbb.yale.edu/chr22) that was confirmed by PCR also covered approximately 4.5% of that chromosome and of which only 4.4% matched in annotated coding exons. While these results are intriguing from a biological point of view, their contribution to the gene finding process was found to be insignificant and therefore excluded from the gene finding algorithm.

Using the reference gene set the optimal combination of gene prediction programs, homology to ESTs, homology to known proteins and protein motifs was tested, at the single nucleotide levels, for predicting correctly a coding nucleotide. The correlation coefficient (CC) was used to merge the sensitivity value (the probability of correctly predicting a coding nucleotide) and specificity value (the probability that a positive example is correct) of each data source (Figure 2). Most predictors had significantly different CC values for large and small genes. Gene prediction programs had a reduced CC value for small genes that was the result of reduced sensitivity, i.e. many exons of small annotated genes were missed by gene prediction programs. Of all gene prediction programs, sgpGene had the highest CC values 0.89 and 0.74 for large and small genes, respectively. Ensembl annotations were nearly perfect for both gene groups, with CC values close to 1. AceView had slightly lower CC values of 0.97 and 0.95 for large and small genes, respectively. Ensembl and AceView both use publicly available mRNA/protein sequences in their annotation process, therefore, their performance was nearly perfect on our set of annotated genes. The contribution of Ensembl and AceView to new non-annotated genes was not clear and therefore they were excluded from the prediction algorithm.

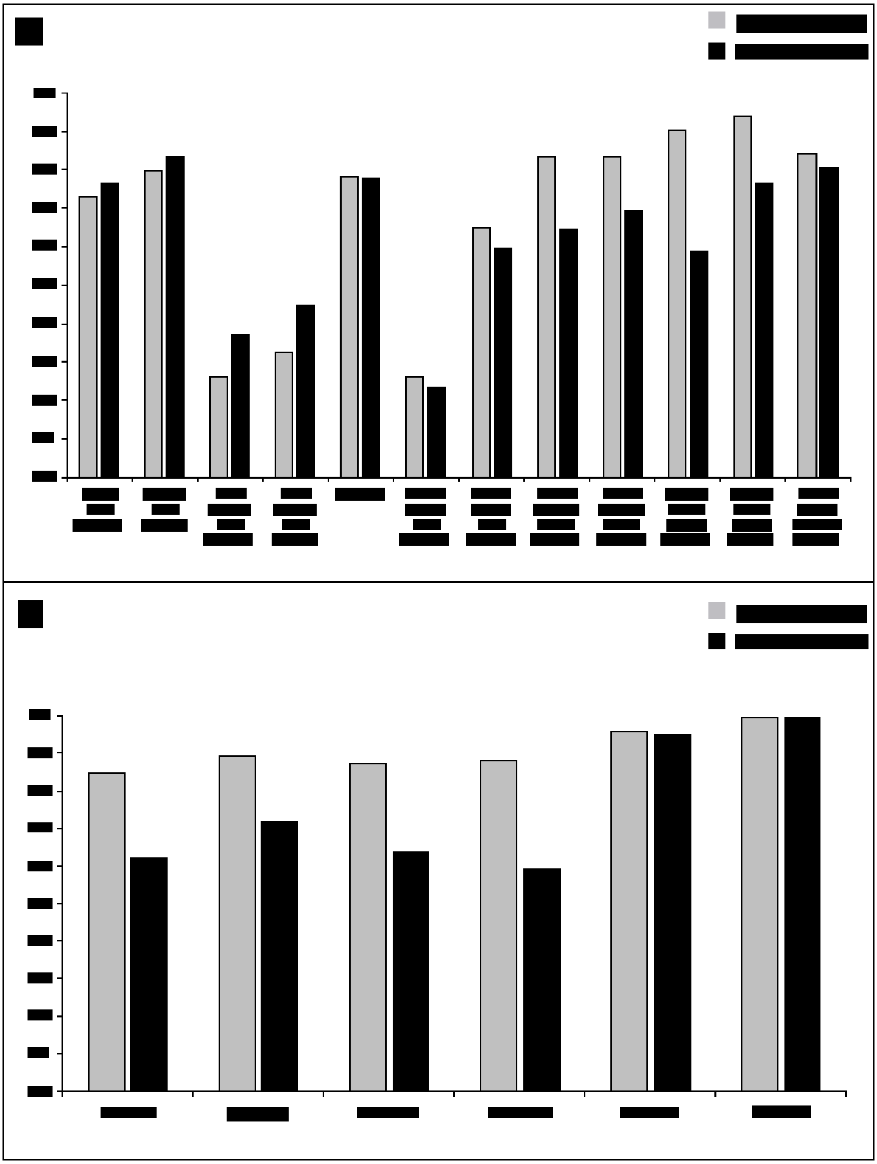


Figure 2: Correlation coefficient values for predicting a protein coding nucleotide for large and small genes. Values were derived from the training set and are presented separately for ESTs and protein data (A) and gene prediction data (B). Nearly all evidences show a reduced CC value for small genes. Binary values refer only to 0 or 1 values while ordinal values refer to the number of matching evidences transformed on a natural logarithmic scale and rounded to an integer. In the case of spliced ESTs, for instance, ordinal values were in the range of 0 to 10.

Using EST data, the binary (True/False) value of spliced ESTs had the highest CC values for both gene groups (Figure 2A). Known proteins that matched a CCP element and were spliced had the highest CC value for large genes. However, a match to any known protein or motif had the highest CC value for small genes. Nearly all nucleotides in coding regions of both gene groups were conserved in dog, rat and mouse. However, the coding regions of small genes were significantly less conserved than large genes in more distant species, namely, fish and chicken (Figure 3).

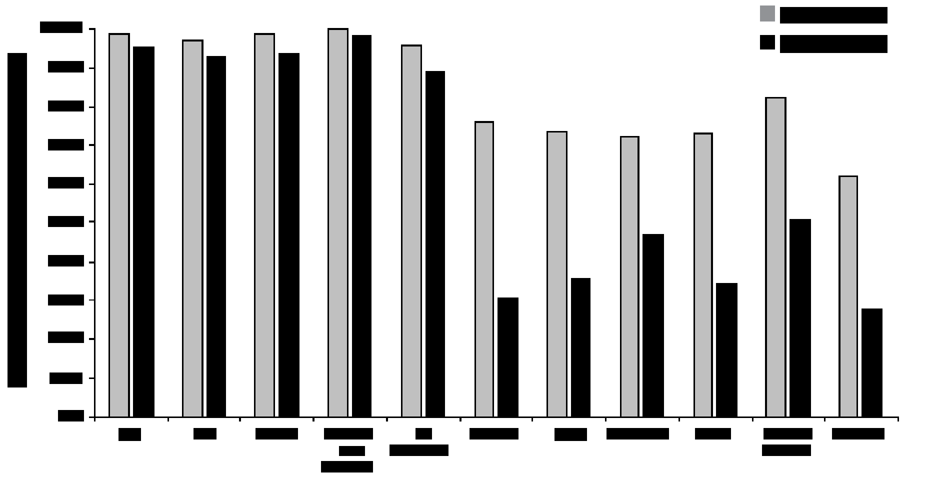


Figure 3: The number of conserved nucleotides between human and other species in large and small genes. At least one mammal includes one of dog, rat and mouse and all mammals include all three. At least one fish includes one of Fugu, Tetraodon and Zebrafish and "All fish" includes all three fish. Small genes were shown to be conserved in mammals similarly to large genes, however, in more distant species small genes were less conserved than large genes.

Several combinations of data sources were used to estimate the probability that a nucleotide is coding for a protein using a logistic regression model, for small genes and large genes separately. The β-coefficients and the prediction equation derived from the model are shown in Table 1. The best predictive models for small and large genes were different in the weights assigned to each data source. In general, the highest β-coefficients were assigned to homology to known proteins and ESTs. However, the contribution of the gene prediction programs to the logistic regression model was significant. Twinscan had the highest weight compared to other gene prediction programs and sgpGene had a relative higher weight for small genes than for large genes. The only conservation data that was found to contribute to the prediction process was PhastCons, a phylogenetic hidden markov model for identifying evolutionarily conserved regions. Interestingly, the identification of protein motifs had very little contribution to the prediction model. Identified motifs had a relatively low weight in the model for large genes and were insignificant and therefore excluded from the small genes model.

A

Data source	β-coefficient
Genscan	1.44
Twinscan	2.42
sgpGene	2.05
PhastCons	0.20
Spliced ESTs	4.28
Non spliced ESTs	0.91
Spliced aligned known proteins	3.88
BlastX alignments	3.19
Constant	-4.56

B

Data source	β-coefficient
Genscan	1.06
Twinscan	3.15
GeneID	0.85
sgpGene	1.74
PhastCons	0.09
Non-spliced ESTs	0.89
Spliced ESTs	2.65
BlastX alignments	1.04
CDD match	0.72
Spliced aligned known proteins	4.00
Constant	-6.01

Table 1: β-coefficients derived by employing a logistic regression analysis to the training set of small (A) and large genes (B). All evidences used except PhastCons were in binary format, i.e. existing (1) or non-existing (0). These coefficients were used to predict small and large genes using the equation $\Pr(\text{coding}|z)=e^z/(1+e^z)$, where $z = \beta_0 + \beta_1 \cdot \text{evidence}_1 + \beta_2 \cdot \text{evidence}_2 + \ldots$ The Cox and Snell and Nagelkerke R square values were 0.7 and 0.94 for small genes and 0.73 and 0.97 for large genes, respectively.

The logistic regression equation was used to score each nucleotide for coding potential. The algorithm presented here maintained the high specificity and increased sensitivity on the test set of small genes, 0.96 and 0.90 respectively (CC value 0.87) (Figure 4). Each nucleotide in CCP elements was scored for the coding potential of small and large genes and exons were generated separately for each gene class by combining adjacent nucleotides with a coding score equal to or larger than 0.5. Adjacent exons were fused into one exon if there was a spliced EST or a matching known protein that bridged the gap between the exons and no other EST or protein supported the existence of two exons. The first percentile of intron sizes in currently annotated genes was 66 bp, therefore only exons which were less than 66 nucleotide apart were allowed to be fused. In many cases, the small genes algorithm produced two overlapping predicted exons on opposite strands. This is probably the result of ESTs, often matching the opposite strand of a gene and conserved elements that are not associated with one of the two strands. Every two overlapping predicted exons on opposite strands were scored as follows: one point per data source for being supported by Acembly, Twinscan, GeneID, Genscan and sgpGene and two points for per data source for being supported by spliced proteins, CDD match, BLASTX match and EnsEmbl. If the difference in score was larger than 1 point and at least 80% of the exon with the lower score was covered by the other it was removed from the prediction set. Genes were generated based on spliced ESTs, matches to known

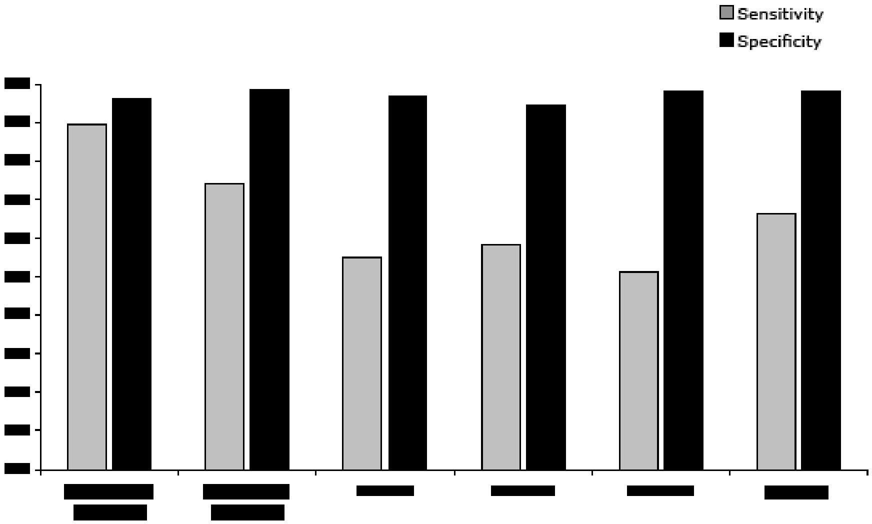


Figure 4: Sensitivity and specificity values for prediction of nucleotides of small genes on a test set of annotated small genes. While many of the evidence show a high specificity (the probability that a positive example is correct) for small genes, they suffer from a reduced sensitivity (the probability of correctly predicting a coding nucleotide), i.e. many of the small genes were missed. Using the logistic regression model the prediction algorithm was able to maintain the high specificity and increase the sensitivity. The model used on a training set of large genes (large genes algorithm) was also shown to have a reduced sensitivity compared to the small genes algorithm.

proteins and gene prediction programs. The optimal combination of data sources was difficult to determine since no negative training set could be compiled with reasonable confidence. A suggested connection between two exons could either agree with current annotations, conflict with current annotations or be classified as unknown. All possible connections between two exons were considered for each data source and combination of data sources and classified into one of the three categories. Most data sources had a 90% agreement with current annotation, less than one percent conflicted and about 10% of the connections between two exons were unknown. Empirical rules were derived that minimize conflicts and maximize agreement with current annotations. A connection between two exons was considered reliable if it was (i) supported by both multiple ESTs and multiple protein alignments, or (ii) supported by one of the two, combined with at least one gene prediction programs, or (iii) supported by at least 4 gene prediction programs. All other possible connections between two exons were considered less reliable. This approach generated 151,000 reliable connections between exons of which 88.5% were supported by current annotations, 0.2% were in conflict with the annotations and 11.3% were unknown. The longest possible gene was generated ignoring possibilities for alternative splicing. Reliable connections between exons were established first and exons with less reliable connections were added if there was no conflict with the other connections (Figure 5).

Genes that were predicted using the small genes algorithm and had a predicted ORF length longer than 300 bp or an in-frame stop codon were discarded. Each predicted exon was classified to one of three locations: overlapping an annotated exon, located within an intron of an annotated gene and located in an intergenic region (between two genes). The small genes prediction algorithm was more sensitive than the large genes prediction algorithm, predicting 42,135 non-annotated exons of small genes, 14 fold more than the large genes algorithm. The average exon size of small predicted non-annotated genes was 82 bp, compared to 97 bp for large genes and 84 bp for annotated small genes. While it is possible that most of these exons are real, it is also possible that many of these are pseudoexons (exons of pseudogenes), non-annotated alternative splicing exons or false positives. Predicted small genes were compared to the 7,395 and 7,358 annotated pseudogenes in the Vega and Yale pseudogenes databases, respectively. Approximately

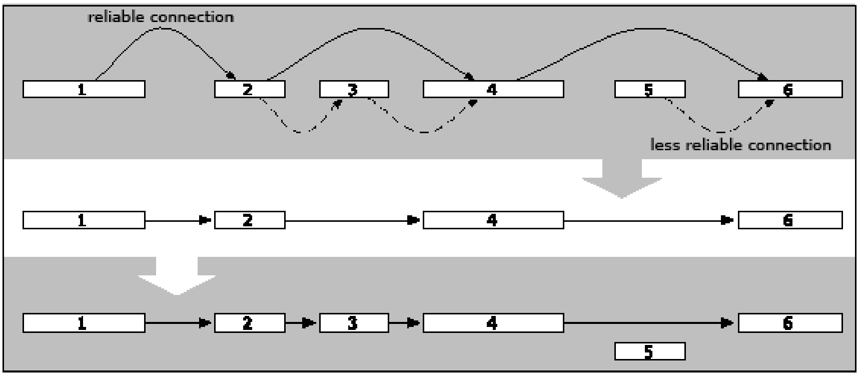


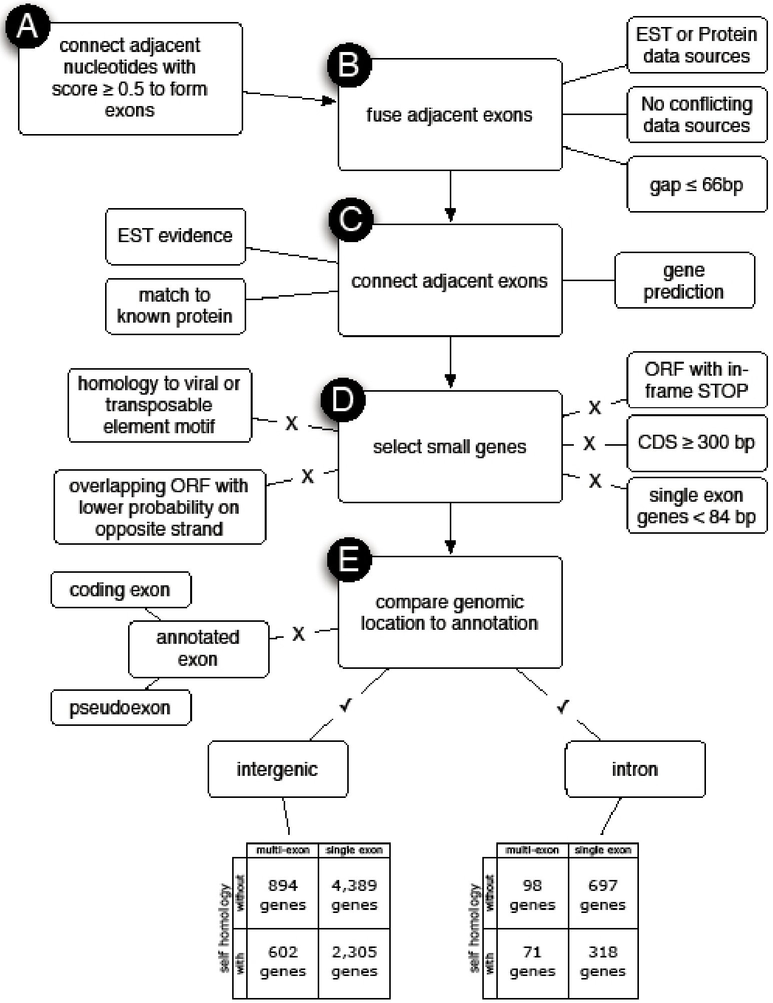
Figure 5: Scheme demonstrating exon connection. Gene prediction programs, ESTs and protein alignments were used to evaluate connections between exons. Reliable connections (solid line) were supported by (i) multiple ESTs and multiple protein alignments or (ii) multiple ESTs or multiple protein alignment with at least one gene prediction program or (iii) at least 4 gene prediction programs. All other connections, such as one gene prediction program or a single EST, were considered less reliable (dashed line). Initially exons with reliable connections were connected to form a gene structure (top section to middle section). Less reliable connections, such as exon 3, were incorporated into the gene structure as long as there was no conflict with the gene structure (middle section to bottom). Exon 5 could not be incorporated into the gene structure since no evidence existed that connected exons 4 to 5. However, exon 5 remained part of this gene region and was not assumed to be a single exon gene.

10% of the Vega annotated pseudogenes and 13% of the Yale pseudogenes had at least one exon that overlapped a predicted exon of a small gene. Those predicted small genes that overlapped annotated pseudogenes were 14.8% of the entire predicted set. Most pseudogenes have homology to genomic sequences elsewhere in the genome, 71% and 79% of the Vega and Yale annotated pseudoexons had such self-homology. In comparison only 19% of the Vega annotated protein coding exons had self-homology. Of all predicted small genes in the genome, 60% of the genes with multiple exons and 66% of the genes with single exons had no self-homology. Predicted small genes with no self-homology were compared to the Vega and Yale pseudogenes annotations and only 1.2% and 1% of the pseudogenes matched a predicted small gene, respectively. The predicted small genes matching annotated pseudogenes represented 3.2% of the predicted small genes in the genome that had no self-homology. Almost 80% of the Vega pseudogenes and all of the Yale pseudogenes were single exon genes. Therefore, predicted multi-exon small genes have a lower probability of being pseudogenes. Only 3.5% and 3.4% of the Vega and Yale pseudogenes match multi-exon predicted small genes, respectively; and 0.3% and 0.1%

matched multi-exon predicted small genes with no self-homology, respectively. In total 15% and 2.7% of the multi-exon predicted small genes with and without self-homology matched annotated pseudogenes, respectively. All predicted small genes that overlapped an annotated pseudogene were excluded from the prediction set.

We employed a series of steps to generate several sets of predicted new small genes with different levels of confidence (Figure 6). Small predicted genes with homology to known proteins and protein motifs were excluded if the protein or motif description included keywords related to viruses (“viral”, “virus”, “virulence”) and the keyword “transposable element”. Approximately 29% of the predicted small genes included such keywords. The remaining genes were classified as follows: (i) located within annotated intron or intergenic regions, (ii) with or without self homology, and (iii) multi-exon or single exon gene. In the case of single exon genes, we selected only genes with a minimum length of 84 bp. This number is based on the observation that the average exon length of annotated small genes was 84 bp and the average largest exon of multi-exon small annotated genes was 131 bp. Assuming that the largest exon can be identified more easily than the other exons and based on the observation that exons of small genes are smaller and more difficult to identify than exons of large genes (chapter 2), single exon genes are more likely to be the largest exon of a small multi-exon gene. The probability that the largest exon would be smaller than the average exon is low. The selection process and results for the different sets are summarized in Figure 6. Seven fold more predicted small genes were located in intergenic regions compared to genes that were predicted within introns of annotated genes. In total 9,374 non-annotated small genes were predicted in the human genome, 1,665 (~18%) were multi-exon small genes and 7,709 (~82%) were single exon small genes. Nearly 65% of the predicted small genes had no self-homology and the subset that is most likely to include real new small genes is a set of 894 multi-exon small genes with no self-homology located in intergenic regions (Figure 7).

Protein motifs from CDD that had homology to predicted small genes were analyzed. While many were annotated as “prediction only” and “Function unknown” several systems were present in larger numbers than others. Examples include: (i) post-translational modification, protein turnover, chaperones; (ii) signal transduction mechanisms; (iii)


A
connect adjacent nucleotides with score ≥ 0.5 to form exons
B
fuse adjacent exons
EST or Protein data sources
No conflicting data sources
gap ≤ 66bp
C
EST evidence
match to known protein
connect adjacent exons
gene prediction
D
homology to viral or transposable element motif
X
overlapping ORF with lower probability on opposite strand
X
select small genes
X
ORF with in-frame STOP
X
CDS ≥ 300 bp
X
single exon genes < 84 bp
E
coding exon
annotated exon
X
pseudoexon
compare genomic location to annotation
✓
✓
intergenic
intron

self homology		multi-exon	single exon
	without	894 genes	4,389 genes
	with	602 genes	2,305 genes

self homology		multi-exon	single exon
	without	98 genes	697 genes
	with	71 genes	318 genes

Figure 6: A scheme illustrating the process of small genes construction and selection. Each nucleotide in CCP elements was scored using the logistic regression model and adjacent nucleotides were connected to form an exon (A). Adjacent exons with a gap smaller than 66 bp and supporting ESTs or protein data sources were fused to form one exon (B). Exons were connected to form a gene structure using ESTs, protein alignments and gene prediction programs (C). Genes with a predicted CDS larger than 300 bp, single exon genes smaller than 84 bp, exons containing an in-frame stop codon or exons containing homology to viral or transposable element motif were excluded (D). Finally the location of each exon was compared to current annotation (E) and all genes with exons matching an annotated pseudoexon or coding exon were excluded. The final set contained intergenic small genes and small genes that occurred within introns of other genes. These were further divided to single and multi-exon small genes, with or without self-homology.

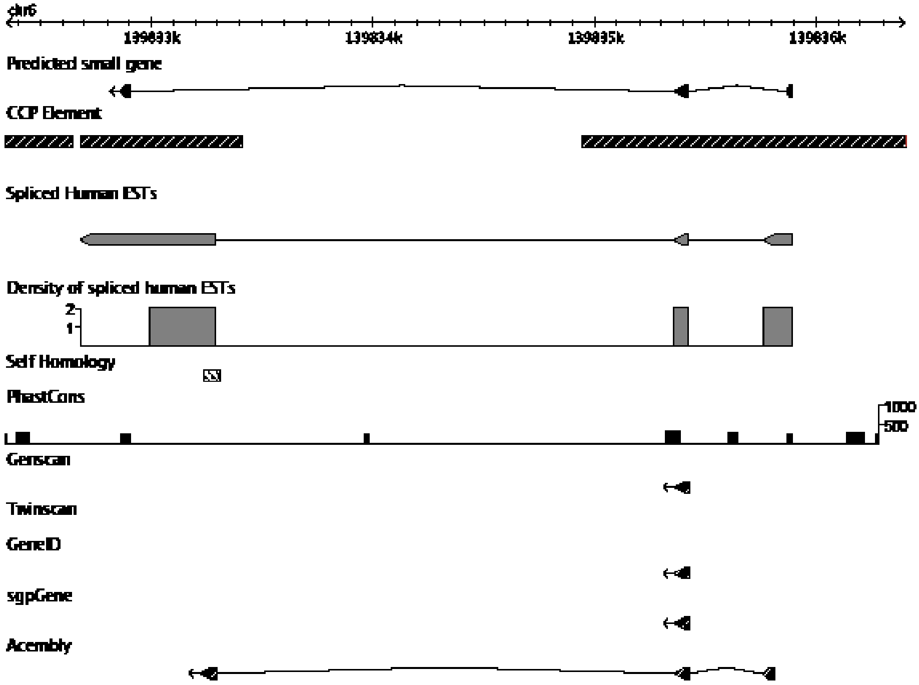


Figure 7: An example of a predicted multi-exon small gene with no self-homology, located in an intergenic region. The prediction is supported by spliced human ESTs, conservation in mammals and one of the three exons is supported by gene prediction programs. This image was generated using the Gbrowse program and only selected data sources are shown for illustration purposes, the full image is available at http://rulai.cshl.edu/smallgenes.

transcription; translation, ribosomal structure and biogenesis; (iv) RNA processing and modification; (v) Energy production and conversion; and (vi) Intracellular trafficking, secretion, and vesicular transport.

DISCUSSION

In the annotation process of the first fully sequenced yeast chromosome Oliver et al. employed a threshold of 100 codons to identify a protein-coding gene (Oliver et al. 1992). This threshold was rationalized by the low likelihood of not encountering a stop codon in a random DNA sequence of 100 codons (Fickett 1994; Fickett 1995). As such, this threshold was initially accepted as a compromise in order not to miss most of the protein coding genes and avoid a large number of non-coding ORFs. However, to this day genome annotators often implement the 100 codons threshold. In their second revision of human chromosome 22, Collins et al. used the 100 codons threshold as the detectable limit for a protein-coding gene (Collins et al. 2003). As a result, chromosome 22 is the only chromosome in the human genome for which no annotated small gene existed in our reference set of 228 genes. Due to the lower probability that a small predicted gene is a real gene in comparison to large predicted genes, genome annotation is systematically biased against finding small genes. The work presented here is an initial attempt to generate an algorithm biased towards identification of small human genes in order to push genome annotation one step further. The algorithm described above starts with a large set of DNA sequences that may contain protein-coding exons and reduces these systematically to a set of predicted small genes. The initial set of CCP elements was shown to be adequate as it included all coding exons of the reference gene set. The most likely CCP elements to contain exons and the exact exon boundaries were selected using gene prediction programs, conserved evolutionary elements and homology to ESTs and to known proteins. Homology to ESTs and known proteins had higher weights than gene prediction programs and spliced protein matches were shown to be more significant than monolithic protein matches. Within gene prediction programs, algorithms that employed additional information in the prediction process, such as Twinscan and sgpGene, performed better than others that used the sequence alone. Despite the lower sensitivity of gene prediction programs in identifying small genes, their input was shown to be valuable by the logistic regression model. On the other hand using conserved evolutionary elements was shown to be of little use, primarily because the level of conservation was high for too many nucleotides that were labeled as negative in our training set. However, using the level of conservation as determined by the phylogenetic hidden markov model PhastCons was shown to be beneficial for both small and large genes though less than other data sources. The algorithm derived here was shown to be reliable by maintaining the high specificity of the individual data sources and increasing the relatively low sensitivity for a test set of annotated small genes. The equivalent large genes algorithm was shown to have a reduced sensitivity, confirming that programs that were trained on a set of large genes are most likely biased against finding small genes. It is reasonable to assume that the majority of gene prediction programs did not include any small genes in their training and testing sets. Less than 3% of the annotated genes are small genes and most developers and users would avoid including small genes in high quality gene sets due to the lower likelihood that they were real genes and the general conception of the 100 codons threshold.

Many more exons were predicted in the human genome using the small genes model compared to the large genes model. Excluding predicted genes based on their size to increase confidence in the predicted set is common practice, however, it reduces the likelihood of identifying real small genes. The high number of predicted exons using the small genes model raises the question as to the extent of small genes in the human genome. The human genome sequencing consortium recently estimated that the number of small human genes is likely to be at most 10% of the current number of annotated genes, i.e. 2000-2500 genes (HGSC 2004).The large number of predicted exons of small genes in intronic and intergenic regions suggest that small genes may exist in greater numbers than estimated. The major difficulties in estimating the number of small genes with a reasonable confidence is (i) the existence of a large number of pseudogenes, most of which contain 1 to 3 pseudoexons and have a short detectable ORF and (ii) the existence of alternatively spliced exons that are not annotated. Using two pseudogenes databases we demonstrated that the algorithm is biased toward avoiding pseudogenes, since more than 85% of the annotated pseudogenes were skipped. However, many more pseudogenes probably exist that are not yet annotated. In order to increase the level of confidence in predicted small genes we searched the human genome for homology to the predicted exons small genes. Only a small fraction of the annotated pseudogenes had no self-homology, which is probably due to the way pseudogenes evolve. Excluding all predicted genes with self-homology is a compromise between increasing the confidence in the prediction and reducing the number of false positive small genes. Given the large number of small predicted genes with no self-homology and the lack of biological knowledge on the extent of small genes, we believe this compromise is acceptable. Most annotated pseudogenes are single exon genes, therefore multi-exon small genes are less likely to be pseudogenes. However, excluding all single exon small genes, especially those with no self-homology, is unreasonable at this point since observations of annotated small genes showed that exons were relatively small but one exon was larger and contained most of the coding sequence (chapter 2). We believe that many of those single exon small genes may be multi-exon small genes with undetected additional exons.

Almost 13% of the predicted small genes were located within introns of annotated genes and the vast majority of these were single exon genes. It is possible that many of these genes are non-annotated alternative splicing exons. Nevertheless, the annotation of the *Drosophila melanogaster* genome showed that genes are present within genes, i.e. genes may often occur one within the intron of another and not systematically one behind the other (Misra et al. 2002). It is therefore more difficult to estimate the extent of small genes that occur within introns of other genes.

Due to the challenges discussed here and in chapter 2, we believe it is difficult to annotate small coding genes with a high degree of confidence based on computational identification alone, even if homology to ESTs and to known proteins exist. To better understand and characterize small genes, many more must first be identified and confirmed biologically. Randomly sequenced mRNA and proteins, currently contribute very little new information, therefore, we suggest a biological approach that uses the predicted sequences as anchors and identifies the entire mRNA sequence using 5′ and 3′ RACE with specific primers (Figure 8). This approach will be implemented by our group, targeting first small genes with a higher likelihood of being correct, i.e. multi-exon intergenic small genes with no self-homology. A set of primers, matching the requirements described in Figure 8, was generated for all predicted small genes and is available to the public at http://rulai.cshl.edu/smallgenes. The task of confirming and correctly annotating small genes is of great importance to the human genome annotation process. Raising the number of identified small genes will allow compiling better datasets to generate more sensitive algorithms and identify new small genes in order to better distinguish true coding genes from pseudogenes.

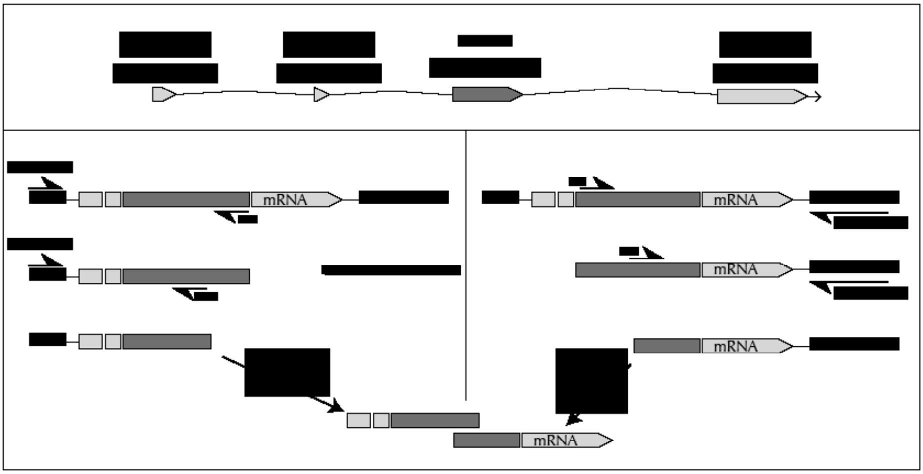


Figure 8: Suggested PCR strategy for using one identified exon as an anchor for "fishing" the entire mRNA sequence. Two separate reactions are carried out, 5′ and 3′ RACE. Initially a universal primer is used (CAP primer for 5′ RACE and polyT for 3′ RACE) with a specific primer, followed by a second reaction using the product of the first as template and another matching specific primer. The second reaction is necessary to increase specificity due to usage of one universal primer. Since exons are short, optimal primers will be difficult to find and may occur partially in other mRNA sequences. However the likelihood of a second primer occurring in the same mRNA is very small. The two products are sequenced and were designed to partially overlap to allow correct assembly of the final mRNA product.

REFERENCES

Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman D.J. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-10.

Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-402.

Ashurst J.L., Chen C.K., Gilbert J.G., Jekosch K., Keenan S., Meidl P., Searle S.M., Stalker J., Storey R., et al. 2005. The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res* **33:** D459-65.

Bairoch A., Apweiler R., Wu C.H., Barker W.C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., et al. 2005. The Universal Protein Resource (UniProt). *Nucleic Acids Res* **33 Database Issue:** D154-9.

Barry C., Fichant G., Kalogeropoulos A. and Quentin Y. 1996. A computer filtering method to drive out tiny genes from the yeast genome. *Yeast* **12:** 1163-78.

Basrai M.A., Hieter P. and Boeke J.D. 1997. Small open reading frames: beautiful needles in the haystack. *Genome Res* **7:** 768-71.

Bateman A., Coin L., Durbin R., Finn R.D., Hollich V., Griffiths-Jones S., Khanna A., Marshall M., Moxon S., et al. 2004. The Pfam protein families database. *Nucleic Acids Res* **32:** D138-41.

Boguski M.S., Lowe T.M. and Tolstoshev C.M. 1993. dbEST--database for "expressed sequence tags". *Nat Genet* **4:** 332-3.

Burge C. and Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268:** 78-94.

Burset M. and Guigo R. 1996. Evaluation of gene structure prediction programs. *Genomics* **34:** 353-67.

Collins J.E., Goward M.E., Cole C.G., Smink L.J., Huckle E.J., Knowles S., Bye J.M., Beare D.M. and Dunham I. 2003. Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Res* **13:** 27-36.

Curwen V., Eyras E., Andrews T.D., Clarke L., Mongin E., Searle S.M. and Clamp M. 2004. The Ensembl automatic gene annotation system. *Genome Res* **14:** 942-50.

Cyster J.G. 1999. Chemokines and cell migration in secondary lymphoid organs. *Science* **286:** 2098-102.

Fickett J.W. 1994. Inferring genes from open reading frames. *Comput Chem* **18:** 203-5.

Fickett J.W. 1995. ORFs and genes: how strong a connection? *J Comput Biol* **2:** 117-23.

Florea L., Hartzell G., Zhang Z., Rubin G.M. and Miller W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* **8:** 967-74.

Ganz T. 2003. Defensins: antimicrobial peptides of innate immunity. *Nat Rev Immunol* **3:** 710-20.

HGSC 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931-45.

Hubbard T., Barker D., Birney E., Cameron G., Chen Y., Clark L., Cox T., Cuff J., Curwen V., et al. 2002. The Ensembl genome database project. *Nucleic Acids Res* **30:** 38-41.

Kapranov P., Cawley S.E., Drenkow J., Bekiranov S., Strausberg R.L., Fodor S.P. and Gingeras T.R. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296:** 916-9.

Karolchik D., Baertsch R., Diekhans M., Furey T.S., Hinrichs A., Lu Y.T., Roskin K.M., Schwartz M., Sugnet C.W., et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res* **31:** 51-4.

Kent W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-64.

Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M. and Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12:** 996-1006.

Korf I., Flicek P., Duan D. and Brent M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17 Suppl 1:** S140-8.

Letunic I., Copley R.R., Schmidt S., Ciccarelli F.D., Doerks T., Schultz J., Ponting C.P. and Bork P. 2004. SMART 4.0: towards genomic data integration. *Nucleic Acids Res* **32:** D142-4.

Marchler-Bauer A. and Bryant S.H. 2004. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* **32:** W327-31.

Marenholz I., Heizmann C.W. and Fritz G. 2004. S100 proteins in mouse and man: from evolution to function and pathology (including an update of the nomenclature). *Biochem Biophys Res Commun* **322:** 1111-22.

Misra S., Crosby M.A., Mungall C.J., Matthews B.B., Campbell K.S., Hradecky P., Huang Y., Kaminker J.S., Millburn G.H., et al. 2002. Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. *Genome Biol* **3:** 83.1-83.22.

Mulder N.J., Apweiler R., Attwood T.K., Bairoch A., Barrell D., Bateman A., Binns D., Biswas M., Bradley P., et al. 2003. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res* **31:** 315-8.

Oliver S.G., van der Aart Q.J., Agostoni-Carbone M.L., Aigle M., Alberghina L., Alexandraki D., Antoine G., Anwar R., Ballesta J.P., et al. 1992. The complete DNA sequence of yeast chromosome III. *Nature* **357:** 38-46.

Parra G., Agarwal P., Abril J.F., Wiehe T., Fickett J.W. and Guigo R. 2003. Comparative gene prediction in human and mouse. *Genome Res* **13:** 108-17.

Parra G., Blanco E. and Guigo R. 2000. GeneID in Drosophila. *Genome Res* **10:** 511-5.

Pruitt K.D., Tatusova T. and Maglott D.R. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33:** D501-4.

Rinn J.L., Euskirchen G., Bertone P., Martone R., Luscombe N.M., Hartman S., Harrison P.M., Nelson F.K., Miller P., et al. 2003. The transcriptional activity of human Chromosome 22. *Genes Dev* **17:** 529-40.

Rudd K.E., Humphery-Smith I., Wasinger V.C. and Bairoch A. 1998. Low molecular weight proteins: a challenge for post-genomic research. *Electrophoresis* **19:** 536-44.

Siepel A., Bejerano G., Pedersen J.S., Hinrichs A.S., Hou M., Rosenbloom K., Clawson H., Spieth J., Hillier L.W., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034-50.

Tatusov R.L., Fedorova N.D., Jackson J.D., Jacobs A.R., Kiryutin B., Koonin E.V., Krylov D.M., Mazumder R., Mekhedov S.L., et al. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4:** 41.

Thierry-Mieg D., Thierry-Mieg J. Identification and functional annotation of cDNA-supported genes in higher organisms, in *http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly.*

Wasinger V.C. and Humphery-Smith I. 1998. Small genes/gene-products in Escherichia coli K-12. *FEMS Microbiol Lett* **169:** 375-82.

Yada T. and Hirosawa M. 1996. Detection of short protein coding regions within the cyanobacterium genome: application of the hidden Markov model. *DNA Res* **3:** 355-61.

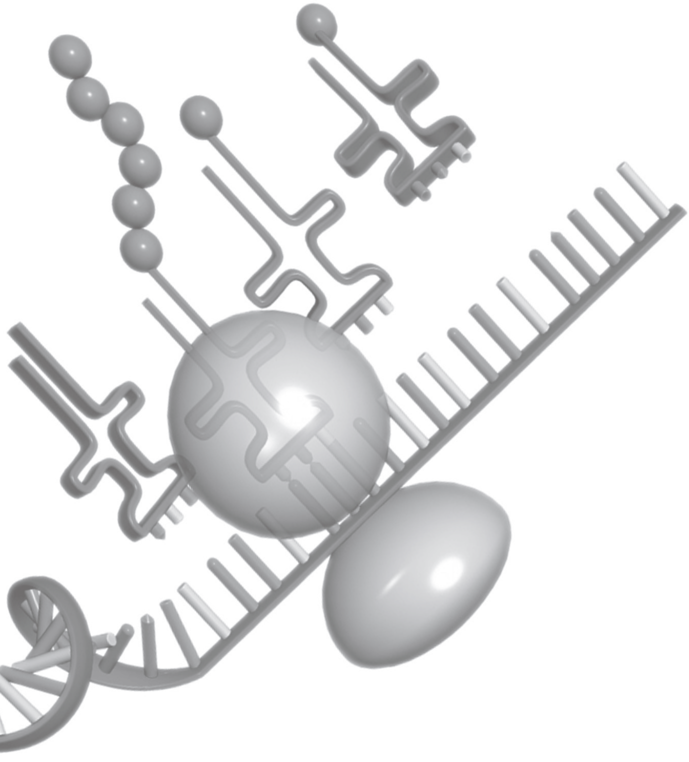
Part II

STRENGTHENING PROTEOMICS FOUNDATIONS

Confronting barriers of
high-throughput protein production

4

Understanding PCR and recombinant protein expression in *Escherichia coli*



INTRODUCTION

Proteomics research is developing in two parallel lines: (i) mass-spectrometry, where proteins are proteolytically cleaved into peptides, separated and analyzed using computer algorithms to determine the sequence identity of the proteins; and (ii) array based methods where proteins or affinity ligands are spotted onto a surface. Protein arrays are used to identify protein-protein interactions, to identify the substrates of protein kinases and or to identify the targets of biologically active small molecules (Bertone and Snyder 2005). Antibody arrays, on the other hand, are used to browse the protein content of a cell or tissue extract (Agaton et al. 2003; Pavlickova et al. 2004).

One of the major bottlenecks of array-based proteomics is the automated production of proteins on a large scale. All current proteomic technologies that allow genome-scale analysis, such as protein arrays and antibody arrays, rely on the production of native proteins or proteins fused to affinity tags. High-throughput (HT) production is the ability to automate protein production on a large scale, e.g. in 96-well format. The major problem is that proteins are very different one from another and unlike DNA, there is a great difficulty to establish one protocol that is suitable for the production of all proteins. Furthermore, the optimal protocol for expression of a specific protein is established along general guidelines and adjusted by trial and error. Even if the optimal conditions were known it would be inconceivable to implement them for each protein when applying HT technology. In this respect any protocol selected for HT production would be a compromise that is not optimal for a specific protein but good enough for the average protein. When establishing a HT platform, in each of the 96-well plates as many as half of the proteins are expected to fail expression, increasing the already high costs of protein production.

Depending on the objectives and aims of the research, proteins can be produced in their native, soluble and correctly folded form or as denatured proteins. In general, producing native and correctly folded proteins is much more difficult. Such proteins are required, for instance, for structural studies that depend on crystallization of the protein. HT production of such proteins has very low success rates of ~10% (Bertone et al. 2001; Goh et al. 2003). Proteins for antibody generation and selection do not require the full protein sequence or that the protein be correctly folded. In this respect protein production for antibody generation and selection is less demanding. The antibodies generated against an unfolded protein fragment will most liekly recognize a linear epitope in the native protein (Barlow et al. 1986).

Several groups attempted to produce proteins using a master protocol in a 96-well format and reported success rates of 60%-80%. Braun et al. expressed 336 randomly selected human cDNAs in *E. coli* and purified successfully 60% under denaturing conditions using His_6 constructs and 50% under non-denaturing conditions using glutathione S-transferase (GST) constructs (Braun et al. 2002). Luan et al. expressed 10,176 Caenorhabditis elegans proteins using a robotic pipeline and observed an overall expression of 50% (15% in soluble form) (Luan et al. 2004). Agaton et al. reported a success rate of 76% for the expression of 142 human proteins in *E. coli* (Agaton et al. 2003). Other groups reported success rates of HT protein expression in *E. coli* in the range of 60%-80% (Christendat et al. 2000; Pizza et al. 2000; Dobrovetsky et al. 2005). Most of these HT methods resolve around the following outline: affinity tags, which are used for protein purification, are selected and a master vector is prepared into which the target sequence is inserted. Target sequences are amplified by PCR using specific primers. The resultant PCR products are incorporated into the master expression vector which is used to transform a bacterial host. The bacteria produces the protein with its fused affinity tags. Finally, the proteins are purified using affinity chromatography. The most significant variation between the different HT protocols is the selected N-terminal or C-terminal fusion proteins. Hammarstrom et al. compared expression and solubility of 32 proteins using 7 different N-terminal fusion proteins, namely, His_6, GST, NusA, ZZ, Gb1, MBP and Thioredoxin (Hammarstrom et al. 2002). Their results showed significant differences in expression and solubility dependent on the fusion protein used. Some fusion proteins perform better than others, but no single construct was able to express all 32 proteins. The overall success rate combining all constructs was 85%. The authors concluded that the key to high overall success was the combination of several fusion possibilities, meaning each protein should be expressed several times each time with another fusion partner. While feasible, such a strategy will increase tremendously the work load and costs of HT production. However, costs and workload could be reduced if we knew *a priori* the most suitable fusion partner for each target protein based on the protein sequence alone. This way each protein could be expressed once in the most suitable construct. Unfortunately, there is currently no reliable prediction algorithm that allows the selection of the optimal fusion partner.

1. HIGH-THROUGHPUT PROTEIN PRODUCTION

In our approach (chapters 6) to HT protein synthesis one main fusion protein was selected, namely, ZZ. The ZZ domain is the tandem repeat dimer of the modified immunoglobulin binding domain of protein A of *Staphylococcus aureus* (Nilsson et al. 1987). An antigenic domain fused to the ZZ moiety is thought to be targeted to antigen presenting cells via IgG molecules present on the surface of these cells (Léonetti et al. 1998) or via anti-CD11c (Wang et al. 2000). An additional advantage of the ZZ domain is that it is a very effective carrier protein for recombinant protein expression in *E. coli* in terms of solubility and stability (Samuelsson et al. 1994; Hammarstrom et al. 2002). Therefore, the ZZ domain is a suitable fusion partner for proteins that will be used to elicit antibodies, both for immunological and protein expression purposes. Two additional tags have been added for purification purposes, namely, His_6 on the N-terminal and streptag (Skerra and Schmidt 2000) on the C-terminal (Figure 1).

Since a fragment of the protein is enough to elicit antibodies against the native protein, one fragment from each protein that corresponds to one exon was selected for expression (see chapter 5 for selection criteria). This allowed the use of human genomic DNA as template for PCR, eliminating the need for obtaining cDNA sequences. Furthermore, shorter sequences are easier to express compared to long sequences (Bertone et al. 2001; Goh et al. 2003), especially if the expressed protein fragment is a small part of a fusion protein. The technical details of primer design for PCR and PCR protocol are presented in chapter 5; the technical details of protein expression in *E. coli* are presented in chapter 6.

The HT protein production procedure can be separated into two independent steps: (i) creating the expression vector; (ii) expression and purification of the recombinant protein in *E. coli* (Figure 1). In the first step, most failure occurs at PCR level, where only ~80% of the fragments were visualized on agarose gel as a single band of expected size (see chapter 5). In the second step most failure occurs at the level of protein expression within

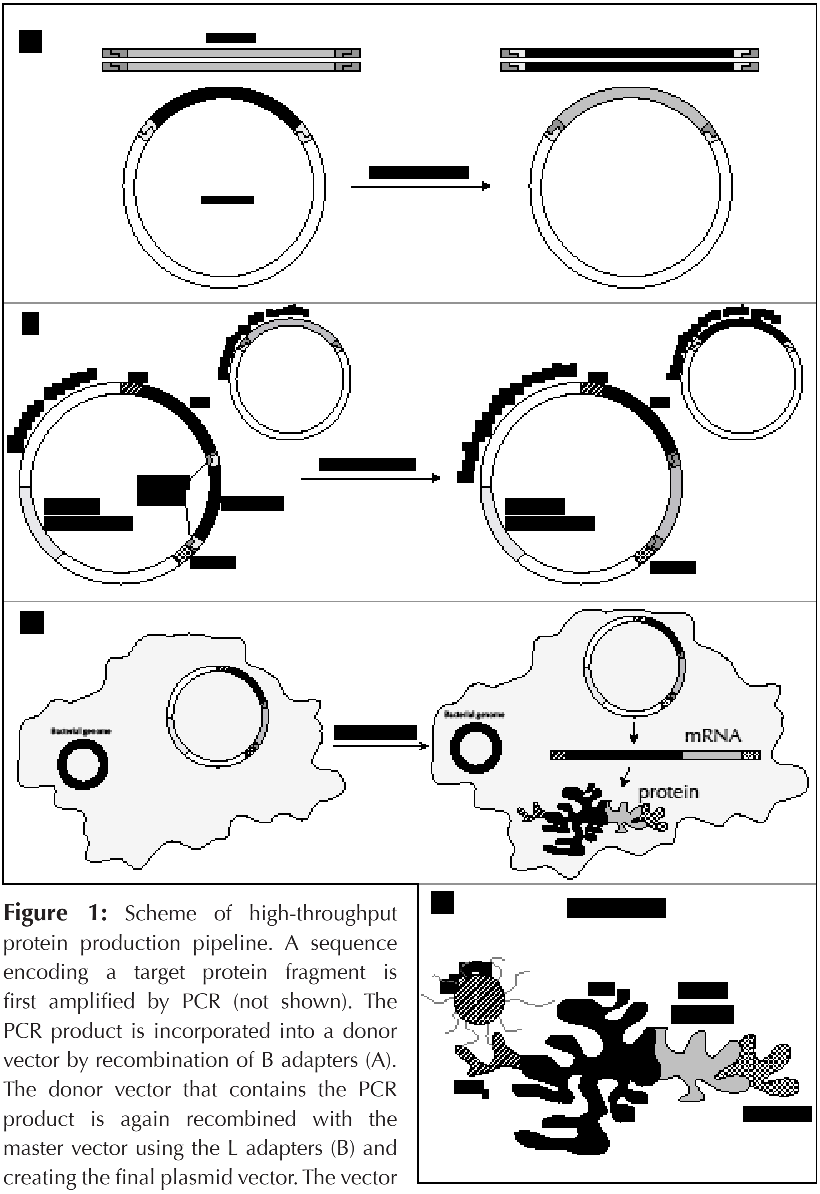


Figure 1: Scheme of high-throughput protein production pipeline. A sequence encoding a target protein fragment is first amplified by PCR (not shown). The PCR product is incorporated into a donor vector by recombination of B adapters (A). The donor vector that contains the PCR product is again recombined with the master vector using the L adapters (B) and creating the final plasmid vector. The vector is transformed into competent E.coli cells. Upon indection the bacteria starts producing the recombinant protein (C). Finally the proteins are purified using Ni-NTA beads that bind the His6 affiny tag. The proteins can be further purified using the streptag.

the bacterial host and only ~50% of the proteins were expressed correctly (see chapter 6). These success rates are in accordance with other studies of HT protein productions (Bertone et al. 2001; Braun et al. 2002; Hammarstrom et al. 2002; Agaton et al. 2003; Luan et al. 2004). In this part of the thesis we attempted to link primary DNA sequence to successful PCR and DNA and protein sequences to successful protein expression. Such a link would make it possible to select DNA sequences that are suitable for our HT PCR protocol and protein sequences that are suitable to our HT protein expression protocol. Instead of using several fusion partners as suggested above, we aim to identify, based on sequence analysis, proteins that are suitable for the HisZZ-streptag construct. In the next section an overview of potential problems of PCR and protein expression are presented that lay the foundation of the DNA and protein sequence analysis performed in chapters 5 and 6.

2. POTENTIAL PROBLEMS IN POLYMERASE CHAIN REACTION (PCR)

PCR is a well-characterized enzymatic reaction for amplification of DNA. Many technical problems may inhibit PCR or reduce reaction yield, such as template dilution, template quality, buffers used, quality of dNTPs, quality of the polymerase enzyme, etc. However, in the HT PCR procedure, as applied in chapter 5, everything except the primers was constant including the PCR template quality and concentration (since genomic human DNA was used as template). Therefore, it was assumed that any problems in PCR were the direct result of the primers or amplicon (the product of PCR).

The various problems that arise during PCR are directly related to DNA hybridization. Hybridization is the process of combining complementary, single-stranded nucleic acids into a double stranded molecule. Nucleotides will bind to their complement under normal conditions, so two perfectly complementary strands will bind to each other readily. Conversely, due to the different geometries of the nucleotides, a single inconsistency between the two strands will prevent them from binding. This process can be reversed by heating the mixture and separating the double stranded molecule resulting in two single stranded molecules, a process called denaturation. PCR is carried out in cycles where each cycle of the reaction requires a brief heat treatment to denature the DNA template and a cool down to anneal the primers and synthesize new complementary strands (Figure 2). Each of these cycles multiplies the amount of DNA, resulting in an exponential amplification of the DNA target sequence.

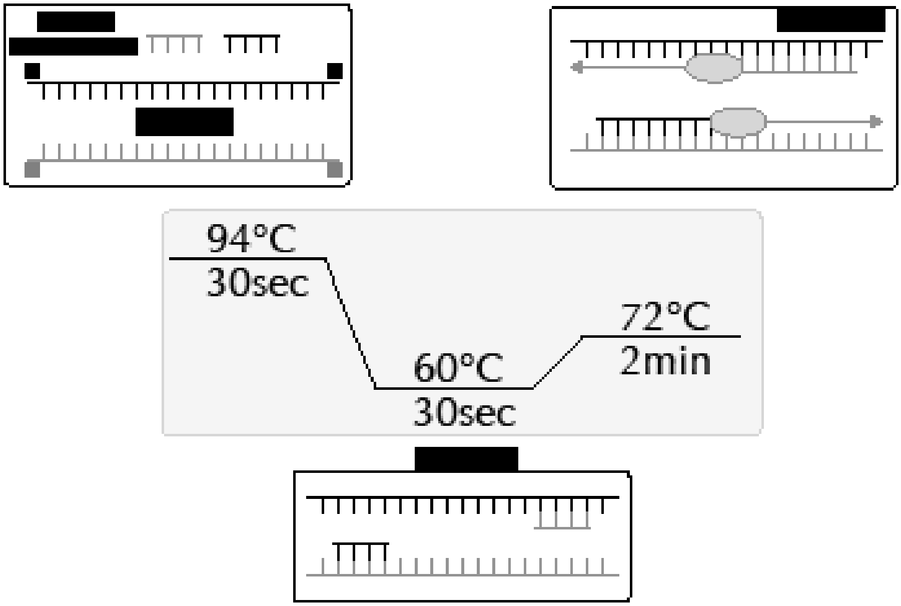


Figure 2: The three thermal steps of a typical PCR cycle. In the first step all DNA double strand molecules are denatured (94°C). In the second step the mixture is cooled down to allow the primers to anneal (60°C). Finally the optimal temperature of the polymerase enzyme is used to allow synthesis of the complementary strand (72°C). This cycle is repeated ~30 times to exponentially amplify the initial DNA template.

2.1 Primers associated problems

PCR primers are short single stranded oligonucleotides that form a stable duplex with a target DNA sequence. The primers allow DNA polymerase to start synthesizing the complementary DNA strand from the 5′ end of the DNA to the 3′ end. It is essential that both primers bind exclusively with the specific target site. Occasionally, depending on the primer sequence, temperature and reaction buffer, primers may bind to a DNA sequence that is not fully complementary (Figure 3). Furthermore, primers are typically

~20 bases long and if a primer forms a stable duplex that is only 10 bp long on the 3′ end of the primer, the polymerase reaction may start (Figure 3C). Such mismatches may occur, however, to amplify a DNA sequence other than the one targeted, the mismatch must be within a reasonable distance from the other primer, since only a sequence that is flanked by two primers on opposite strands can be amplified exponentially.

The dissociation of a DNA double helix is often referred to as melting because it occurs abruptly once a certain temperature has been reached. The melting temperature (T_m) is defined as the temperature at which 50% of the DNA molecules form a stable double helix and the other 50% are in the form of single strand molecules. One of the most important attributes of primers for PCR is their T_m. The T_m of a primer is determined by the primer sequence and the concentration of Mg^{+2} in the buffers used for PCR. At high

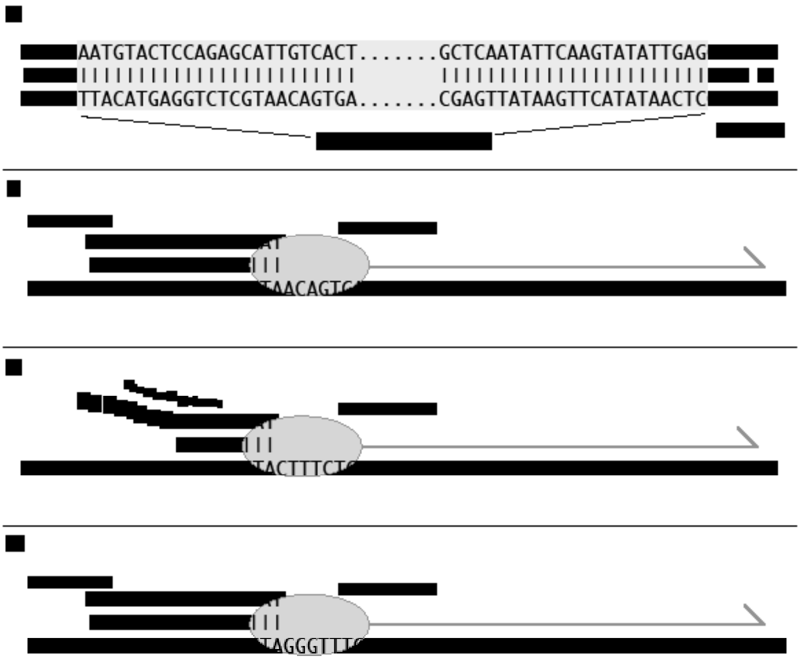

AATGTACTCCAGAGCATTGTCACT.......GCTCAATATTCAAGTATATTGAG
TTACATGAGGTCTCGTAACAGTGA.......CGAGTTATAAGTTCATATAACTC
TAACAGTG
TACTTTCT
TAGGGTT

Figure 3: Possible mismatches in annealing of PCR primers. Ideally, a primer should form a stable duplex with the a sequence that is a perfect match (B). However, depending on the temperature and buffers mismatches may occur. A primer may form a stable duplex only on the 3′ end allowing for the reaction to start (C) or bind along the entire length of the primer with some mismaches in between (D).

temperature fewer primers anneal to the target DNA site and mismatches are tolerated. As the temperature decreases more primers anneal to the target DNA site and mismatches are less tolerated; however, lowering the temperature too much may cause hybridization of the DNA template or PCR product. The T_m of the primers is correlated with the GC/AT ratio and typically this ratio should be similar to or higher than that of the amplified template (Rychlik et al. 1993). The nearest neighbor thermodynamic parameters (see chapter 5) have been shown to produce reliable predictions of T_m for short oligonucleotides, such as PCR primers (Breslauer et al. 1986; Freier et al. 1986; Wu et al. 1991; Sugimoto et al. 1996).

Since PCR is carried out from the 5′ end to the 3′ end, stable binding of the primer to the template on the 3′ end is more important than binding on the 5′ end (Figure 3C). To further increase primer specificity, the stability of the 3′ end of the primer should be lowered so that primers would need the entire length for forming a stable duplex. Furthermore, sequence complexity on the 3′ end should be as high as possible, avoiding any kind of repeats. A primer pair could also form a stable duplex one with the other, instead of binding to their intended target site, or form a self-complementary hairpin loop rendering the primer unavailable for PCR (Figure 4) (Rychlik et al. 1993). These are design flaws that must be taken into account while designing primers and should be avoided if possible.

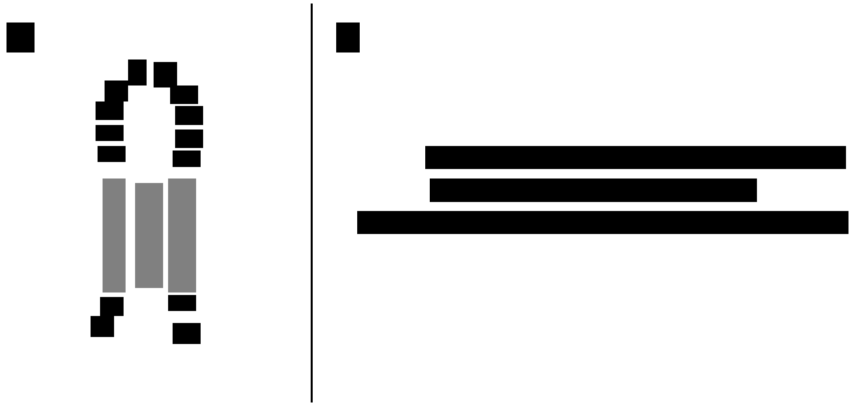


Figure 4: A self-complementary sequence of a primer that may render the primer unavailable for PCR (A) and two primers that form a primer-dimer may also reduce PCR yield.

2.2 PCR template consideration

Potential problems regarding the target sequence are often overlooked. The template has to be fully denatured, available for the primers to bind and must not form any kind of secondary structure that could disrupt the progress of the DNA polymerase. Rychlik et al. showed clearly that the T_m of the PCR product affected the efficiency of the reaction (Rychlik et al. 1990). The authors defined the optimal PCR temperature (T_a^{OPT}) as:

$$T_a^{OPT} = 0.3xT_m^{Primer} + 0.7xT_m^{Product} - 14.9$$

Their equation suggest that the T_m of the PCR product has a more significant affect on PCR efficiency than the T_m of the primers. It should be noted that the algorithms for determining product T_m are not as reliable as those predicting primer T_m. The nearest neighbor method is only reliable for short oligonucleotides, for longer sequences prediction algorithms rely on GC content and thus are less accurate. Despite problems associated with the PCR product, in many cases the target sequence is needed and no selection is possible. In cases where the user can select part of the template, the T_m of the product should be taken into account.

3. POTENTIAL PROBLEMS IN PROTEIN EXPRESSION

In this section the process of protein expression is examined, assuming the target plasmid is inside the bacteria and the only variable element in the process is the DNA sequence encoding the protein. All parameters that may affect protein expression, such as promoters used, bacterial growth medium and growth temperature, are assumed to be constant. This situation is common when establishing a protocol for HT production as opposed to optimizing a protocol for expression of an individual protein. The expression process is divided into three stages: (i) plasmid and transcription (ii) mRNA and translation (iii) the target protein and its effects. Expression can fail in any of the three stages and usually only the end result is known.

3.1 Plasmid DNA vector and transcription efficiency

This section deals with the stability of the plasmid that was introduced into the bacteria and the process of transcription in which an mRNA molecule is produced. A plasmid encoding the protein must maintain a sufficient copy number within the bacteria and be available for transcription. Bacterial selection is usually done by means of an antibiotic resistance gene. Only bacteria carrying the plasmid will survive in antibiotic containing growth media and therefore any living bacteria carries the plasmid and is able to transcribe its antibiotic resistance mRNA and produce the antibiotic resistance protein in sufficient quantity. However, the bacteria degrades the antibiotics and over time plasmid-free bacteria may grow. Furthermore, bacteria with a very low plasmid copy number may survive since a small quantity of antibiotic resistance protein is required for survival while a much larger quantity is needed for the target recombinant protein. In a pipeline approach the plasmid vector is the same except for the DNA sequence encoding the protein. Plasmids are typically 5-6Mb long and the effect of a relatively small (typically a few hundred bp long) variable DNA sequence encoding the protein on plasmid stability or transcription is not clear. A DNA sequence that is relatively short compared to the plasmid vector may hypothetically prompt recombination, leading to non-functional vectors or affect the secondary structure of the plasmid and inhibit transcription of the protein encoding part. Both options are currently hard to quantify by sequence analysis. However, other properties of DNA that can be calculated from the primary sequence may have an indirect effect on plasmid stability. Those include DNA sequence complexity, DNA bending and DNA curvature. Low complexity sequences, such as repetitions of CAG or CTG triplets, was shown to form intranstrand hairpin loops with combinations of normal and mismatched base pairs that easily rearrange (Hartenstine et al. 2000; Petruska et al. 1996). Bending and curvature are local conformational micropolymorphism of DNA in which the original DNA structure is only distorted but not extensively modified. Such conformational micropolymorphism have been hypothesized to act as 'environmental sensors' whose conformational transitions act as regulatory signals (Gabrielian et al. 1997). Sequence complexity was further linked to DNA curvature, showing that low complexity segments were preferentially located in close proximity to the highly curved sequences (Gabrielian and Bolshoy 1999). Despite potential problems that may cause plasmid instability or low transcription efficiency, once a plasmid vector has been designed and was shown to be working as expected for a few proteins it is unlikely that failure would occur at this stage for other proteins.

3.2 mRNA stability and translation efficiency

The two mRNA characteristics that are of major importance in successful protein expression are mRNA half-life and translation efficiency. The production and decay of mRNA is of high importance in the control of gene expression and is strictly regulated by various cellular components. A particular mRNA can only be translated a limited number of times before it will be functionally inactivated. In bacteria mRNA half-life can vary from 40 second to 20 minutes (Kushner et al. 1996). Decay of mRNA can become a problem when the recombinant mRNA has a very fast decay. In such a case only very few proteins will be produced on each recombinant mRNA template resulting in an overall low yield. It is now well established that mRNA decay is not a default process, in which an array of nonspecific nucleases degrades indiscriminately based on target size or ribosome protection of the substrate. Rather, like transcription, RNA processing, and translation, mRNA decay is a precise process dependent on a variety of specific cis-acting sequences and trans-acting factors (Jacobson and Peltz 1996). In the case of protein expression in bacteria, the 5′ and 3′ ends of the mRNA sequence are usually identical for all recombinant proteins, therefore, a vector can be designed to have 5′ and 3′ secondary structures or motifs that are known to stabilize bacterial mRNA. The primary question is: how does the variable DNA fragment that encodes the protein affect mRNA decay? So far, a number of factors, which affect mRNA stability in bacteria, have been identified. Carrier and Keasling summarized these into five categories (Carrier and Keasling 1997): Secondary structures, translation effects, nucleotide sequence effects, transcription effects and cellular growth effects. In the following section each of these categories is discussed.

3.2.1 Secondary Structure

RNA is a single stranded molecule that is unstable by design. One of the important attributes of RNA is the ability to increase stability by forming a secondary structure (Figure 5). While this attribute is essential for functional RNA molecules, it can cause many problems in heterologous protein expression. The mRNA template must not form a secondary structure that interferes with ribosomal binding or ribosomal movement. On the other hand it should be stable enough to resist breakdown by RNase. The rate-limiting step in mRNA decay is usually an initial endonucleolytic cleavage (Apirion 1973; Ehretsmann et al. 1992; Grunberg-Manago 1999) and an mRNA with no secondary structure will be degraded immediately. A recombinant mRNA for protein expression is usually designed to have a stable 5′ and 3′ structures to resist degradation. These structures are constant for all recombinant proteins and only an internal part of the mRNA, which encodes the protein, is variable. However, this part may be prone to cleavage by endonucleases resulting in a short mRNA half-life and eventually a low protein yield. Unfortunately, predicting mRNA structure by sequence alone, on a large scale, is still unreliable (Eddy 2004). Even for a known mRNA structure it would be very difficult to estimate the ability to resist endonucleases or the level of interference with ribosome movement.

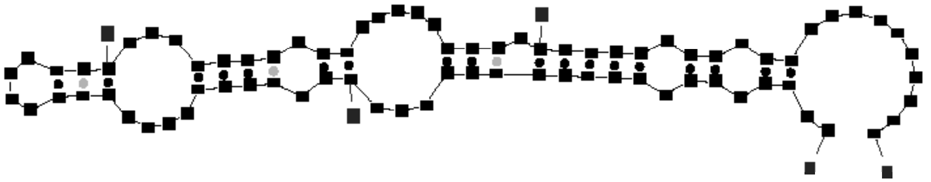


Figure 5: Predicted mRNA secondary structure of a fragment from human stress activated protein kinase 4. Prediction was performed using mfold (Zuker et al. 2003) with default parameters. The black dots signify a correct match between nucleotides and the gray dots signify a non-optimal but week link between two nucleotides.

3.2.2 Translation effects

Translation effects relate to the theory of ribosome protection that originated from observations that the number of ribosomes on a transcript influences the stability of mRNA (Belasco and Brawerman 1993). Increased ribosome loading, i.e. many ribosomes translate the mRNA simultaneously, increases the frequency with which cleavage sites will be covered and shielded from the endonucleases. Therefore, any parameter that affects ribosome movement could affect mRNA stability by causing a gap of unprotected mRNA region.

3.2.2.1 codon usage

The genetic code is degenerated, therefore most amino acids can be encoded by more than one codon (Crick 1966). As DNA sequences of numerous genes were determined, it became apparent that the usage of alternative codons for the same amino acid was neither uniform nor random (for a review see (Ikemura 1985). Furthermore, it was noted that a part of this non-random usage is species specific (Grantham et al. 1981) and within a specie there is considerable heterogeneity between genes (Sharp and Li 1987) (Figure 6). In some genes the codon usage seems random (no codon bias), while in others there is a clear preference to use specific codons to encode the same amino acid (codon bias). In the two best-studied organisms, namely *E. coli* and *Saccharomyces cerevisiae*, there is a clear correlation between degree of codon bias and level of gene expression (Bennetzen and Hall 1982; Gouy and Gautier 1982). Major codons are codons that occur in highly expressed genes, whereas minor or rare codons tend to be in genes expressed at low levels. This division, although intriguing from an evolutionary point of view, could be a problem when the goal is to make large quantities of a heterologous synthetic protein (Kane 1995). The preferred codon usage of *E. coli* or of any other species was shown to correlate with the content of its tRNAs (Ikemura 1985) (Figure 6B).Therefore, the high-level expression

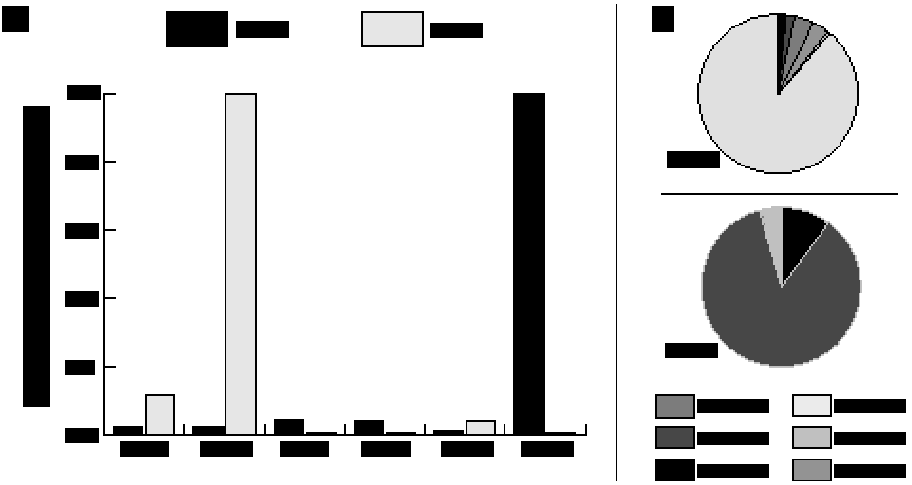


Figure 6: Codon adaptation index values for six codons encoding the amino acid leucine according to Sharp and Li 1987 (A). The most frequently used codon in a set of highly expressed genes is set to 1 and the other codons are scored accordingly. The bias is clearly seen for each of the species. In *E. coli* there is a strong preference to use CUG when encoding leucine, while in yeast (*Saccharomyces cerevisiae*) the preference is to use UUG. The tRNA abundance for each codon is correlated with the codon usage (B).

of a cloned heterologous gene with a codon usage that is not matched with the tRNA population of the host may place demands which the host has difficulties to satisfy. As a result, ribosomes translating the heterologous mRNA may stall at positions calling for a tRNA that is either rare by itself, or largely deacylated because of the heavier than normal drain of its amino acid into protein (Kurland and Gallant 1996). These codons, which are thus involved in the imbalance between demand and supply in the protein synthetic apparatus of the host, are often referred to as hungry codons. Depending on the number of rare codons and their location both expression levels may be reduced and frameshifting may occur (Kane 1995; Barak et al. 1996). Kane et al. reported a translational abberation in the expression of a cloned bovine gene in *E. coli* (Kane et al. 1992). They observed a protein variant that was missing two amino acids, most likely due to in-frame hoping over a disfavored arginine AGG codon. Eliminating disfavored codons abolished the production of the protein variant. Seetherm et al. showed that the use of the AGA codon for arginine, which is rare in *E. coli*, resulted in a significant substitution of lysine (encoded by AAA and AAG) for arginine in rapidly translated IGF-I (Seetharam et al. 1988). Using CGT, a preferred arginine codon for *E. coli*, this mistranslation was abolished. Sorensen et al. showed that rare codons may decrease the rate of translation by up to sixfold (Sorensen et al. 1989). Del Tito et al. were able to increase the yield of heterologous proteins containing frequent isoleucine AUA codons by simultaneous expression of the cloned *ileX* gene for the tRNA that reads that codons (Del Tito et al. 1995). Despite these examples, Swartz et al. suggested that codon usage by itself has, at most, a modest effect on recombinant protein production (Swartz et al. 1996). This is also supported by the results of Ernst and Kawashima who expressed eight α-factor somatomedin-c with codon bias indices in the range of 0.1-0.8 and observed no correlation to expression levels in *E. coli* and *S. cervisiae* (Ernst and Kawashima 1988). The codon adaptation index (CAI), as devised by Sharp and Li, was based on the analysis of 26 highly expressed genes from *E. coli* and attempted to quantify codon bias (Sharp and Li 1987). Their assumption was that highly expressed genes use optimal codons and by analyzing the codon usage of those genes they indexed each codon relative to the frequency of usage of its alternative codons (Figure 6A). The CAI has been revisited in 2003 using a much larger set of genes and was found to be reliable (Jansen et al. 2003). Codon substitution for genes with a low CAI (not optimal) is not applicable for HT systems. Lower amounts of protein produced may be tolerated, however, frameshifts must be avoided. Frameshifts in *E. coli* can be shown only with the most rare arginine codons, namely AGA and AGG (Kane et al. 1992) and such genes should be carefully analyzed. Special strain of *E. coli* have been specifically designed for protein expression and carry extra tRNA genes that recognize rare codons in *E. coli*. One such example is the BL21-CodonPlus-RIL (Stratagene) that contains extra tRNAs for AGA and AGG (arginine), AUA (isoleucine) and CUA (leucine). These bacterial strains were developed to enable expression of proteins encoded by genes rich in GC. Overall, codon usage should be taken into account but the effect on the bacterial host is expected to be moderate.

3.2.2.2 Amino acids usage and starvation

Over-expression of a protein could quickly drain the amino acids available for protein synthesis causing a starvation-like effect. In conditions of nutritional stress, a global change in cellular metabolism is initiated, a phenomenon termed 'stringent response' (Cashel et al. 1996). Briefly, lack of amino acids changes the ratio of aminoacylated tRNA to free tRNA. When a free tRNA is encountered at the A-site of the 50S ribosome, protein synthesis is stalled, resulting in an idling reaction in which ribosome-bound RelA is activated to synthesize guanosine 3′, 5′-bispyrophosphate (ppGpp) (Chatterji and Ojha 2001). The accumulation of ppGpp appears to be the trigger, which initiates the 'stringent response' and is an important link between nutritional stress and bacterial adaptation (Cashel et al. 1996). Over-expression of recombinant proteins with an amino acid content significantly different than the "average" *E. coli* protein might lead to a situation similar to nutritional stress and thus to the stringent response. Harcum and Bentley 1999 compared the stringent response, heat-shock response and recombinant protein over-expression (Harcum and Bentley 1999). They concluded that the over-expression of a recombinant protein appears to induce a stress response that is overlapping but not identical to either the stringent response or heat-shock response. Effects of such a stress response may result in reduced accumulation of mRNA (Ishihama and Creighton 1999); negative influence on DNA replication, recombination and repair (Autret et al. 1999; McGlynn and Lloyd 2000); decreased bacterial growth; increased protein degradation (Maurizi 1992; Kuroda et al. 1997; Kuroda et al. 1999) and elevated protease activity that continues long after the stress is removed (Harcum and Bentley 1999). These events are more likely to take place when a strong promoter is used that forces fast production of the recombinant protein leading to amino acid starvation (Swartz and Neidhardt 1996). This is usually the case in heterologous protein over-expression.

A kinetic model to predict intracellular amino acid shortage during recombinant protein over-expression in *E. coli* was suggested (Harcum 2002). According to this model, even an average protein, designed to match the chemical amino acid content of *E. coli*, is predicted to cause a shortage in amino acids during over-expression. This prediction showed that the over-expression of recombinant proteins is a very demanding process by itself. Therefore, even a small deviation in amino acid content might trigger a stringent response in *E. coli*. As the deviation increases the bacterial growth rate is likely to be reduced and the protease activity increased.

Lobry and Gautier analyzed the amino acid content of 999 native *E. coli* proteins and showed a bias in amino acid usage for highly expressed genes (Lobry and Gautier 1994). They further showed that genes with high CAI values tend to use amino acids who's major tRNA are abundant. Their explanation was a straightforward adaptation of what was visible at the codon level; highly expressed genes reduced the diversity of codon choices to increase translation efficiency. By analogy, proteins encoded by highly expressed genes use a reduced diversity of amino acid choices to increase translation efficiency. The CAI is commonly used to determine whether a native *E. coli* gene is highly expressed and to estimate whether a heterologous protein is likely to be well expressed in *E. coli*. The CAI must be used cautiously when estimating probability of heterologous expression since other factors must be considered. For instance, if a gene with a high CAI value (optimal codon usage) is over-expressed, it may increase the translation efficiency, but if the diversity of amino acids usage is significantly different than the average *E. coli* protein, the end result might be a very low yield due to amino acids shortage. An over-expressed protein can accumulate to 30% and more of the total protein content of the cell, therefore, it poses a significant burden on the host. On the one hand a high CAI is beneficial for the ribosomal machinery of *E. coli* but on the other hand increasing the translational efficiency increases the burden on the host, depleting the amino acid pools and leading to a stringent response.

Ramirez and Bentley tested the effect of glucose, IPTG (inducer) and phenylalanine (rate-limiting precursor) on *E. coli* cells during over-expression of CAT (chloramphenicol acetyltransferase) (Ramírez and Bentley 1995). Phenylalanine is present in CAT in significantly higher levels compared to the "average" *E. coli* protein. Their results showed that the progressive addition of phenylalanine along with IPTG resulted in improved CAT synthesis and stability during exponential growth. They were able to show correlation between demand and consumption of phenylalanine. Furthermore, they showed a significantly lower proteolytic activity upon phenylalanine feeding.

3.2.2.3 Low complexity

A protein with low complexity sequence is a protein that contains repetitive elements. Low complexity can occur both at DNA level and at protein level, however, a low complexity amino acid sequence does not necessarily indicate low complexity at the nucleotide level and vice versa. Many proteins that are prone to aggregation have low complexity amino acid sequence. In Huntington's disease, for instance, a polyglutamine repeat causes aggregation that leads to a pathological state (Ross et al. 2003); another example where low complexity proteins aggregate is spider-web proteins which can instantly turn from soluble proteins to strings stronger than steel (Kenney et al. 2002). Expression of proteins with low complexity in *E. coli* may cause aggregation and inclusion body formation more easily than other proteins. However, a more significant problem may arise due to bias in amino acid composition. Production of a protein with repeat sequences may drain much faster amino acids used in the low complexity region into the produced protein, leading to starvation and stringent response as discussed above. Low complexity has also been implicated in intrinsically disordered regions of proteins (discussed below).

3.2.3 nucleotide sequence effects

Nucleotides sequence effects refer to the recognition of specific RNase cleavage sites and poly(A) tail additions. Cleavage sites are thought to be the rate limiting step in mRNA stability, once cleaved the mRNA fragments are degraded quickly. Avoiding mRNA cleavage sites and poly(A) signals is usually done in the design phase of the master plasmid vector. However, a cleavage site or a sequence that is prone to RNase cleavage may be introduced in the variable sequence encoding the specific protein.

RNase recognition sites are difficult to predict since they usually follow a weak consensus sequence or a combination of sequence and structure. RNase E recognition sites, for instance, were shown to cleave preferably in AU rich regions (Arraiano et al. 1997). However, the density of RNase E cleavage sites was not predictive of mRNA half-life of native *E. coli* proteins (Bernstein et al. 2002). The process of mRNA degradation is an ongoing research effort that is still not fully understood, therefore, it cannot be analyzed efficiently in the prediction of successful protein expression.

3.2.4 transcription effects

Under most conditions, the rate of transcription initiation and elongation does not appear to have a detectable effect on mRNA stability in *E. coli* (Carrier and Keasling 1997). However, several studies have shown that increasing the rate of transcription elongation for strong promoters resulted in less stable mRNA (Chow and Dennis 1994; Makarova et al. 1995). Iost and Dreyfus suggested that this effect may be related to shielding of mRNA by ribosomes since the stability of mRNA depended upon the simultaneous transcription and translation (Iost and Dreyfus 1995). Transcription effects related to transcription initiation are not likely to be changed in HT production since most of the mRNA sequence is constant for all vectors except the internal variable part encoding the protein. Elongation rate, on the other hand, may change at the variable part, however, there is currently no reliable way to predict it from sequence.

3.2.5 cellular growth effects

Growth conditions refer to the medium and temperature in which bacteria are grown. Growth conditions can dramatically affect mRNA stability (Cannons and Pendleton 1994;

Albertson and Nyström 1994), however, those condition remain constant when using a pipeline approach. Conditions should be adjusted in the process of protocol optimization to match the requirements of the majority of the proteins being expressed.

3.3 The target protein and its effects

Once a target protein has been produced inside the bacterial host it can interact with other native proteins, interfere with cellular pathways and disturb the delicate homeostasis of the cell. This situation may occur more frequently when a native and functional protein is being produced in its soluble form. When expressing protein fragments or recombinant proteins it is unlikely that these proteins are functional, however, interactions with other cellular components of the host is possible. This situation can be monitored by following the bacterial growth rate. If the bacteria grow slower than expected or unable to grow at all it is possible that the protein produced is toxic. In the following section several aspects of post protein production are discussed.

3.3.1 Inclusion bodies

A protein can be produced in the bacteria as a soluble protein or in inclusion bodies. Inclusion bodies are insoluble aggregates that typically contain 80-95% recombinant protein and other contaminations such as outer membrane proteins, ribosomal components, small amount of phospholipids and nucleic acids that likely absorb upon cell lysis (Valax and Georgiou 1993). Inclusion bodies are formed from structure folding intermediates (King et al. 1996; Speed et al. 1995; Speed et al. 1996). In general proteins that fail to reach a native conformation rapidly or to interact with folding modulators are more likely to aggregate in inclusion bodies due to misfolding (Baneyx and Mujacic 2004). The likelihood of misfolding is increased by the presence of strong promoters and high inducer (e.g. IPTG) concentrations that are routinely used in recombinant protein production. Nevertheless, inclusion bodies are desired in several cases. By forcing a toxic recombinant protein to form inclusion bodies, large quantities of protein can still be produced without affecting the host (Lee et al. 2000). Furthermore, proteins in inclusion bodies are protected to some extent from proteolysis and thus large amount of protein can be recovered (Haught et al. 1998; Lee et al. 1998; Zhang et al. 1998). Inclusion bodies are also easily separated from cell lysates. However, the protein is recovered in a non-active form and must be dissolved and renaturated to obtain a biologically active protein. This is the main disadvantage of producing proteins in inclusion bodies.

Many different parameters affect the rate of protein production and folding. However, here we focus on the contribution of a specific protein sequence to these properties, assuming all other parameters remain constant, e.g. vector used for expression, bacterial growth conditions, etc. Mitraki et al. showed that a specific mutation could interfere with the folding process of a protein and trap it in the inclusion body state (Mitraki et al. 1991). For each protein a large number of conformations are sterically available during the folding process (Ramachandran and Sasisekharan 1968). Many of the aggregates formed are liable to be kinetic traps corresponding to local energetic minima. Mitraki et al. assumed that a class of sites and sequences stabilize the native state of a protein. Their role could be blocking off-pathway interaction or destabilizing incorrect conformation (Mitraki et al. 1991). Unfortunately, no one was able to characterize those elements yet. Nevertheless, the secondary structure of a protein and its amino acid sequence can provide some clues to the rate and complexity of the native fold. (i) Hydrophobicity and charge play a role in the formation of inclusion bodies by affecting the folding rate. The bacterial cytoplasm is an aqueous environment and hydrophobic regions of the protein form spontaneously a hydrophobic core. Regions of the proteins that are positively or negatively charged are more soluble and therefore may slow down the folding rate of the protein and may require the help of folding modulators. This statement is also supported by the observation that aliphatic amino acids and amino acids that increase thermostability (Gromiha et al. 1999) occur in higher frequency in soluble over-expressed proteins than in proteins that form inclusion bodies (Idicula-Thomas and Balaji 2005). Several algorithms have been devised that link amino acids sequence and protein hydrophobicity (Kyte and Doolittle 1982; Bull and Breese 1974; Guy 1985; Engelman et al. 1986) Proteins from the SPINE database, a database for tracking results of high-throughput protein expression for structural studies (Bertone et al. 2001), that were observed to be soluble upon over-expression were shown to correlate with several factors including the minimum GES hydrophobicity score (Engelman et al. 1986) over all amino acids in a 20 residue sequence window. (ii) The size of a protein domain can also affect folding rate. In general, small single domain proteins have fast folding kinetics, whereas large multidomain and over-expressed recombinant proteins often require the assistance of folding modulators. (iii) Secondary structure elements, such as β-sheet or α-helix have different folding kinetics. A β-hairpin was shown to fold at a rate of about 30 times slower than the rate of α-helix formation (Muñoz et al. 1997). A study of β-lactamase inclusion bodies by Raman spectroscopy showed that proteins in inclusion bodies have a lesser amount of α-helix and a larger amount of β-sheet than native proteins (Przybycien et al. 1994). A computational study of 145 proteins, also suggested that inclusion body-forming proteins have a higher sheet propensity, whereas soluble proteins have a higher helix propensity (Idicula-Thomas and Balaji 2005). (iv) Intrinsic protein disorder. Intrinsic protein disorder refers to segments or to whole proteins that fail to fold completely on their own (Romero et al. 2001). These segments are characterized by low sequence complexity, with amino acid compositional bias and high-predicted flexibility (Dunker et al. 1998; Garner et al. 1998; Romero et al. 2001). Characteristics of intrinsically disordered proteins include: a similar structure to proteins denatured by urea or guanidine; susceptibility to enzymatic degradation; failure to spontaneously fold correctly; and higher abundance in bacteria than in eukaryotes. To locate these natively unfolded sequences in nature, they have developed the computer program PONDR (Predictors Of Natural Disordered Regions), a collection of various predictors that function from primary sequence information (Romero et al. 1997; Romero et al. 1998; Romero et al. 2001). Their results suggest that nature is rich in natively disordered protein (Dunker et al. 2000). Other tools have been developed to identify disordered regions, namely, NORSP (Liu and Rost 2003), DisEMBL (Lindin et al. 2003a), DISOPRED (Ward et al. 2004), GlobPlot (Linding et al. 2003b) and FoldIndex (Prilusky et al. 2005). Those tools are mostly based on hydrophobicity and protein charge calculations.

Despite all these sequence attributes that are likely to affect inclusion bodies formation, there is currently no reliable algorithm to predict the solubility of a protein upon over-expression. Inclusion bodies formation is related to the folding process and folding kinetics, a process that is intensively studied by the scientific community with limited *in silico* prediction success.

3.3.2 Post translational modifications

In many proteins modifications are required once the protein is produced. Eukaryotic proteins, for instance, often contain di-sulfide bonds that are formed between two cysteine amino acids. In the wild type *E. coli* the formation of intra- or intermolecular disulfides is not possible in the reducing cytoplasm, but only in the cell envelope. The inability to form the di-slufide bonds in disulfide bond-rich proteins will often result in the aggregation of the protein products (Baneyx and Mujacic 2004). However, in the case of HT production for affinity ligands this issue is of little concern. Proteins are not expected to fold correctly and aggregation is favorable, as outlines above.

3.3.3 Proteolysis

Proteolytic degradation of protein products causes many problems in the bioprocess of recombinant protein production. Many solutions have been suggested from host modifications to protease inhibitors (Rozkov and Enfors 2004). However, it is clear that a true non-proteolytic cell cannot exist since proteolysis is essential for many metabolic processes. Proteolysis guarantees that abnormal polypeptides do not accumulate within the cell and allows amino acid recycling. Targets for degradation include prematurely terminated polypeptides, proteolytically vulnerable folding intermediates that are kinetically trapped off-pathway, and partially folded proteins that have failed to reach a native conformation (Baneyx and Mujacic 2004). The susceptibility of a protein to cleavage or degradation by a protease is determined by (i) the flexibility of the protein structure; (ii) the extent to which cleavage-prone sequences are exposed; and (iii) the nature of the local interactions made by sidechains of its flanking residues. Proteins of higher structural stability, such as hyperthermophile proteins, commonly show higher resistance to proteolysis (Mukherjee and Guptasarma 2005). “Stable” proteins are those, which have half-lives of several hours.

Proteins in inclusion bodies are largely protected from the proteolysis machinery, however, soluble proteins are not and their half-life after production is dependent on their resistance to bacterial proteases. Previously a protein instability index has been developed that predicted *in vivo* half-life of a protein based on its amino acids sequence. It was shown that proteins that have an *in vivo* half-life of less than 5 hours and more than 16 hours had an instability index higher than 40 and lower than 40, respectively (Guruprasad et al. 1990).

In the next two chapters hundreds of PCRs and synthesized proteins were subject to sequence analysis, using a large number of different bioinformatics tools and algorithms, in an attempt to develop approaches to enhance the efficiency of HT protein expression. Several statistical and machine learning approaches were used to correlate successful expression and attributes derived from primary sequences, putting the suggested theoretical problems discussed here to the test.

REFERENCES

Agaton C., Galli J., Höidén Guthenberg I., Janzon L., Hansson M., Asplund A., Brundell E., Lindberg S., Ruthberg I., et al. 2003. Affinity proteomics for systematic protein profiling of chromosome 21 gene products in human tissues. *Mol Cell Proteomics* **2:** 405-14.

Albertson N.H. and Nyström T. 1994. Effects of starvation for exogenous carbon on functional mRNA stability and rate of peptide chain elongation in Escherichia coli. *FEMS Microbiol Lett* **117:** 181-7.

Apirion D. 1973. Degradation of RNA in Escherichia coli. A hypothesis. *Mol Gen Genet* **122:** 313-22.

Arraiano C.M., Cruz A.A. and Kushner S.R. 1997. Analysis of the in vivo decay of the Escherichia coli dicistronic pyrF-orfF transcript: evidence for multiple degradation pathways. *J Mol Biol* **268:** 261-72.

Autret S., Levine A., Vannier F., Fujita Y. and Seror S.J. 1999. The replication checkpoint control in Bacillus subtilis: identification of a novel RTP-binding sequence essential for the replication fork arrest after induction of the stringent response. *Mol Microbiol* **31:** 1665-79.

Baneyx F. and Mujacic M. 2004. Recombinant protein folding and misfolding in Escherichia coli. *Nat Biotechnol* **22:** 1399-408.

Barak Z., Lindsley D. and Gallant J. 1996. On the mechanism of leftward frameshifting at several hungry codons. *J Mol Biol* **256:** 676-84.

Barlow D.J., Edwards M.S. and Thornton J.M. 1986. Continuous and discontinuous protein antigenic determinants. *Nature* **322:** 747-8.

Belasco J.G. and Brawerman G. 1993. *Control of messenger RNA stability.* Academic Press, San Diego, CA.

Bennetzen J.L. and Hall B.D. 1982. Codon selection in yeast. *J Biol Chem* **257:** 3026-31.

Bernstein J.A., Khodursky A.B., Lin P.H., Lin-Chao S. and Cohen S.N. 2002. Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc Natl Acad Sci U S A* **99:** 9697-702.

Bertone P., Kluger Y., Lan N., Zheng D., Christendat D., Yee A., Edwards A.M., Arrowsmith C.H., Montelione G.T. and Gerstein M. 2001. SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res* **29:** 2884-98.

Bertone P. and Snyder M. 2005. Advances in functional protein microarray technology. *FEBS J* **272:** 5400-11.

Braun P., Hu Y., Shen B., Halleck A., Koundinya M., Harlow E. and LaBaer J. 2002. Proteome-scale purification of human proteins from bacteria. *Proc Natl Acad Sci U S A* **99:** 2654-9.

Breslauer K.J., Frank R., Blöcker H. and Marky L.A. 1986. Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci U S A* **83:** 3746-50.

Bull H.B. and Breese K. 1974. Surface tension of amino acid solutions: a hydrophobicity scale of the amino acid residues. *Arch Biochem Biophys* **161:** 665-70.

Cannons A.C. and Pendleton L.C. 1994. Possible role for mRNA stability in the ammonium-controlled regulation of nitrate reductase expression. *Biochem J* **297 (Pt 3):** 561-5.

Carrier T.A. and Keasling J.D. 1997. Controlling messenger RNA stability in bacteria: strategies for engineering gene expression. *Biotechnol Prog* **13:** 699-708.

Cashel M., Gentry D.R., Hernandez V.J. and Vinella D. 1996. The stringent response, In *Escherichia coli and Salmonella* (ed. Neidhardt), pp. 1458-96. ASM, Washington DC.

Chatterji D. and Ojha A.K. 2001. Revisiting the stringent response, ppGpp and starvation signaling. *Curr Opin Microbiol* **4:** 160-5.

Chow J. and Dennis P.P. 1994. Coupling between mRNA synthesis and mRNA stability in Escherichia coli. *Mol Microbiol* **11:** 919-31.

Christendat D., Yee A., Dharamsi A., Kluger Y., Gerstein M., Arrowsmith C.H. and Edwards A.M. 2000. Structural proteomics: prospects for high throughput sample preparation. *Prog Biophys Mol Biol* **73:** 339-45.

Crick F.H. 1966. The genetic code. 3. *Sci Am* **215:** 55-60 passim.

Del Tito B.J., Ward J.M., Hodgson J., Gershater C.J., Edwards H., Wysocki L.A., Watson F.A., Sathe G. and Kane J.F. 1995. Effects of a minor isoleucyl tRNA on heterologous protein translation in Escherichia coli. *J Bacteriol* **177:** 7086-91.

Dobrovetsky E., Lu M.L., Andorn-Broza R., Khutoreskaya G., Bray J.E., Savchenko A., Arrowsmith C.H., Edwards A.M. and Koth C.M. 2005. High-throughput production of prokaryotic membrane proteins. *J Struct Funct Genomics* **6:** 33-50.

Dunker A.K., Garner E., Guilliot S., Romero P., Albrecht K., Hart J., Obradovic Z., Kissinger C. and Villafranca J.E. 1998. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac Symp Biocomput* 473-84.

Dunker A.K., Obradovic Z., Romero P., Garner E.C. and Brown C.J. 2000. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* **11:** 161-71.

Eddy S.R. 2004. How do RNA folding algorithms work? *Nat Biotechnol* **22:** 1457-8.

Ehretsmann C.P., Carpousis A.J. and Krisch H.M. 1992. Specificity of Escherichia coli endoribonuclease RNase E: in vivo and in vitro analysis of mutants in a bacteriophage T4 mRNA processing site. *Genes Dev* **6:** 149-59.

Engelman D.M., Steitz T.A. and Goldman A. 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem* **15:** 321-53.

Ernst J.F. and Kawashima E. 1988. Variations in codon usage are not correlated with heterologous gene expression in *Saccharomyces cerevisiae* and *Escherichia coli*. *Journal of Biotechnology* **7:** 1-9.

Freier S.M., Kierzek R., Jaeger J.A., Sugimoto N., Caruthers M.H., Neilson T. and Turner D.H. 1986. Improved free-energy parameters for predictions of RNA duplex stability. *Proc Natl Acad Sci U S A* **83:** 9373-7.

Gabrielian A. and Bolshoy A. 1999. Sequence complexity and DNA curvature. *Comput Chem* **23:** 263-74.

Gabrielian A., Vlahovicek K. and Pongor S. 1997. Distribution of sequence-dependent curvature in genomic DNA sequences. *FEBS Lett* **406:** 69-74.

Garner E., Cannon P., Romero P., Obradovic Z. and Dunker A.K. 1998. Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite Differing Structural Characterization. *Genome Inform Ser Workshop Genome Inform* **9:** 201-13.

Goh C.S., Lan N., Echols N., Douglas S.M., Milburn D., Bertone P., Xiao R., Ma L.C., Zheng D., et al. 2003. SPINE 2: a system for collaborative structural proteomics within a federated database framework. *Nucleic Acids Res* **31:** 2833-8.

Gouy M. and Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res* **10:** 7055-74.

Grantham R., Gautier C., Gouy M., Jacobzone M. and Mercier R. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res* **9:** r43-74.

Gromiha M.M., Oobatake M. and Sarai A. 1999. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophys Chem* **82:** 51-67.

Grunberg-Manago M. 1999. Messenger RNA stability and its role in control of gene expression in bacteria and phages. *Annu Rev Genet* **33:** 193-227.

Guruprasad K., Reddy B.V. and Pandit M.W. 1990. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng* **4:** 155-61.

Guy H.R. 1985. Amino acid side-chain partition energies and distribution of residues in soluble proteins. *Biophys J* **47:** 61-70.

Hammarstrom M., Hellgren N., van Den Berg S., Berglund H. and Hard T. 2002. Rapid screening for improved solubility of small human proteins produced as fusion proteins in Escherichia coli. *Protein Sci* **11:** 313-21.

Harcum S.W. 2002. Structured model to predict intracellular amino acid shortages during recombinant protein overexpression in E. coli. *J Biotechnol* **93:** 189-202.

Harcum S.W. and Bentley W.E. 1999. Heat-shock and stringent responses have overlapping protease activity in Escherichia coli. Implications for heterologous protein yield. *Appl Biochem Biotechnol* **80:** 23-37.

Hartenstine M.J., Goodman M.F. and Petruska J. 2000. Base stacking and even/odd behavior of hairpin loops in DNA triplet repeat slippage and expansion with DNA polymerase. *J Biol Chem* **275:** 18382-90.

Haught C., Davis G.D., Subramanian R., Jackson K.W. and Harrison R.G. 1998. Recombinant production and purification of novel antisense antimicrobial peptide in Escherichia coli. *Biotechnol Bioeng* **57:** 55-61.

Idicula-Thomas S. and Balaji P.V. 2005. Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli. *Protein Sci* **14:** 582-92.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* **2:** 13-34.

Iost I. and Dreyfus M. 1995. The stability of Escherichia coli lacZ mRNA depends upon the simultaneity of its synthesis and translation. *EMBO J* **14:** 3252-61.

Ishihama A. and Creighton T.C. 1999. Stringent control, In *Encyclopedia of Molecular Biology* pp. 2451-5. John Wiley & Sons,

Jacobson A. and Peltz S.W. 1996. Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells. *Annu Rev Biochem* **65:** 693-739.

Jansen R., Bussemaker H.J. and Gerstein M. 2003. Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic Acids Res* **31:** 2242-51.

Kane J.F. 1995. Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. *Curr Opin Biotechnol* **6:** 494-500.

Kane J.F., Violand B.N., Curran D.F., Staten N.R., Duffin K.L. and Bogosian G. 1992. Novel in-frame two codon translational hop during synthesis of bovine placental lactogen in a recombinant strain of Escherichia coli. *Nucleic Acids Res* **20:** 6707-12.

Kenney J.M., Knight D., Wise M.J. and Vollrath F. 2002. Amyloidogenic nature of spider silk. *Eur J Biochem* **269:** 4159-63.

King J., Haase-Pettingell C., Robinson A.S., Speed M. and Mitraki A. 1996. Thermolabile folding intermediates: inclusion body precursors and chaperonin substrates. *FASEB J* **10:** 57-66.

Kurland C. and Gallant J. 1996. Errors of heterologous protein expression. *Curr Opin Biotechnol* **7:** 489-93.

Kuroda A., Murphy H., Cashel M. and Kornberg A. 1997. Guanosine tetra- and pentaphosphate promote accumulation of inorganic polyphosphate in Escherichia coli. *J Biol Chem* **272:** 21240-3.

Kuroda A., Tanaka S., Ikeda T., Kato J., Takiguchi N. and Ohtake H. 1999. Inorganic polyphosphate kinase is required to stimulate protein degradation and for adaptation to amino acid starvation in Escherichia coli. *Proc Natl Acad Sci U S A* **96:** 14264-9.

Kushner S.R., Neidhardt R., Curtiss III J.L., Ingraham E.C.C., Lin K.B., Low B., Magasanik W.S., Reznikoff M., Riley M. and Schaechter H.E. 1996. mRNA Decay, In *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology* pp. 849-60. ASM Press, Washington D.C.

Kyte J. and Doolittle R.F. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157:** 105-32.

Lee J.H., Kim J.H., Hwang S.W., Lee W.J., Yoon H.K., Lee H.S. and Hong S.S. 2000. High-level expression of antimicrobial peptide mediated by a fusion partner reinforcing formation of inclusion bodies. *Biochem Biophys Res Commun* **277:** 575-80.

Lee J.H., Minn I., Park C.B. and Kim S.C. 1998. Acidic peptide-mediated expression of the antimicrobial peptide buforin II as tandem repeats in Escherichia coli. *Protein Expr Purif* **12:** 53-60.

Léonetti M., Thai R., Cotton J., Leroy S., Drevet P., Ducancel F., Boulain J.C. and Ménez A. 1998. Increasing immunogenicity of antigens fused to Ig-binding proteins by cell surface targeting. *J Immunol* **160:** 3820-7.

Linding R., Jensen L.J., Diella F., Bork P., Gibson T.J. and Russell R.B. 2003a. Protein disorder prediction: implications for structural proteomics. *Structure* **11:** 1453-9.

Linding R., Russell R.B., Neduva V. and Gibson T.J. 2003b. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res* **31:** 3701-8.

Liu J. and Rost B. 2003. NORSp: Predictions of long regions without regular secondary structure. *Nucleic Acids Res* **31:** 3833-5.

Lobry J.R. and Gautier C. 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. *Nucleic Acids Res* **22:** 3174-80.

Luan C.H., Qiu S., Finley J.B., Carson M., Gray R.J., Huang W., Johnson D., Tsao J., Reboul J., et al. 2004. High-throughput expression of C. elegans proteins. *Genome Res* **14:** 2102-10.

Makarova O.V., Makarov E.M., Sousa R. and Dreyfus M. 1995. Transcribing of Escherichia coli genes with mutant T7 RNA polymerases: stability of lacZ mRNA inversely correlates with polymerase speed. *Proc Natl Acad Sci U S A* **92:** 12250-4.

Maurizi M.R. 1992. Proteases and protein degradation in Escherichia coli. *Experientia* **48:** 178-201.

McGlynn P. and Lloyd R.G. 2000. Modulation of RNA polymerase by (p)ppGpp reveals a RecG-dependent mechanism for replication fork progression. *Cell* **101:** 35-45.

Mitraki A., Fane B., Haase-Pettingell C., Sturtevant J. and King J. 1991. Global suppression of protein folding defects and inclusion body formation. *Science* **253:** 54-8.

Mukherjee S. and Guptasarma P. 2005. Direct proteolysis-based purification of an overexpressed hyperthermophile protein from Escherichia coli lysate: a novel exploitation of the link between structural stability and proteolytic resistance. *Protein Expr Purif* **40:** 71-6.

Muñoz V., Thompson P.A., Hofrichter J. and Eaton W.A. 1997. Folding dynamics and mechanism of β-hairpin formation. *Nature* **390:** 196-9.

Nilsson B., Moks T., Jansson B., Abrahmsén L., Elmblad A., Holmgren E., Henrichson C., Jones T.A. and Uhlén M. 1987. A synthetic IgG-binding domain based on staphylococcal protein A. *Protein Eng* **1:** 107-13.

Petruska J., Arnheim N. and Goodman M.F. 1996. Stability of intrastrand hairpin structures formed by the CAG/CTG class of DNA triplet repeats associated with neurological diseases. *Nucleic Acids Res* **24:** 1992-8.

Pavlickova P., Schneider E.M. and Hug H. 2004. Advances in recombinant antibody microarrays. *Clin Chim Acta* **343:** 17-35.

Pizza M., Scarlato V., Masignani V., Giuliani M.M., Aricò B., Comanducci M., Jennings G.T., Baldi L., Bartolini E., et al. 2000. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* **287:** 1816-20.

Prilusky J., Felder C.E., Zeev-Ben-Mordehai T., Rydberg E.H., Man O., Beckmann J.S., Silman I. and Sussman J.L. 2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **21:** 3435-8.

Przybycien T.M., Dunn J.P., Valax P. and Georgiou G. 1994. Secondary structure characterization of β-lactamase inclusion bodies. *Protein Eng* **7:** 131-6.

Ramachandran G.N. and Sasisekharan V. 1968. Conformation of polypeptides and proteins. *Adv Protein Chem* **23:** 283-438.

Ramírez D.M. and Bentley W.E. 1995. Fed-Batch Feeding and Induction Policies that Improve Foreign Synthesis and Stability by Avoiding Stress Responses. *Biotechnology and bioengineering* **47:** 596-608.

Romero P., Obradovic Z., Kissinger C.R., Villafranca J.E., Garner E., Guilliot S. and Dunker A.K. 1998. Thousands of proteins likely to have long disordered regions. *Pac Symp Biocomput* 437-48.

Romero P., Obradovic Z., Li X., Garner E.C., Brown C.J. and Dunker A.K. 2001. Sequence complexity of disordered protein. *Proteins* **42:** 38-48.

Romero, Obradovic and Dunker K. 1997. Sequence Data Analysis for Long Disordered Regions Prediction in the Calcineurin Family. *Genome Inform Ser Workshop Genome Inform* **8:** 110-24.

Ross C.A., Poirier M.A., Wanker E.E. and Amzel M. 2003. Polyglutamine fibrillogenesis: the pathway unfolds. *Proc Natl Acad Sci U S A* **100:** 1-3.

Rozkov A. and Enfors S.O. 2004. Analysis and control of proteolysis of recombinant proteins in Escherichia coli. *Adv Biochem Eng Biotechnol* **89:** 163-95.

Rychlik W., Spencer W.J. and Rhoads R.E. 1990. Optimization of the annealing temperature for DNA amplification in vitro. *Nucleic Acids Res.* **18:** 6409-12.

Rychlik W., White B.A. and Walker J.M. 1993. Selection of primers for polymerase chain reaction, In *PCR protocols: current methods and applications* pp. 31-40. Humana Press, Totowa.

Samuelsson E., Moks T., Nilsson B. and Uhlen M. 1994. Enhanced in vitro refolding of insulin-like growth factor I using a solubilizing fusion partner. *Biochemistry* **33:** 4207-11.

Seetharam R., Heeren R.A., Wong E.Y., Braford S.R., Klein B.K., Aykent S., Kotts C.E., Mathis K.J., Bishop B.F., et al. 1988. Mistranslation in IGF-1 during over-expression of the protein in Escherichia coli using a synthetic gene containing low frequency codons. *Biochem Biophys Res Commun* **155:** 518-23.

Sharp P.M. and Li W.H. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15:** 1281-95.

Skerra A. and Schmidt T.G. 2000. Use of the Strep-Tag and streptavidin for detection and purification of recombinant proteins. *Methods Enzymol* **326:** 271-304.

Sorensen M.A., Kurland C.G. and Pedersen S. 1989. Codon usage determines translation rate in Escherichia coli. *J Mol Biol* **207:** 365-77.

Speed M.A., Wang D.I. and King J. 1996. Specific aggregation of partially folded polypeptide chains: the molecular basis of inclusion body composition. *Nat Biotechnol* **14:** 1283-7.

Speed M.A., Wang D.I. and King J. 1995. Multimeric intermediates in the pathway to the aggregated inclusion body state for P22 tailspike polypeptide chains. *Protein Sci* **4:** 900-8.

Sugimoto N., Nakano S., Yoneyama M. and Honda K. 1996. Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res* **24:** 4501-5.

Swartz J.R., Neidhardt R., Curtiss III J.L., Ingraham E.C.C., Lin K.B., Low B., Magasanik W.S., Reznikoff M., Riley M. and Schaechter H.E. 1996. E.coli Recombinant DNA Technology, In *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology* pp. 1693-711. ASM Press, Washington D.C.

Swartz R.S. 1996. Escherichia coli recombinant DNA technology, In *Escherichia coli and Salmonella* (ed. Neidhardt), pp. 1693-711. ASM, Washington DC.

Valax P. and Georgiou G. 1993. Molecular characterization of β-lactamase inclusion bodies produced in Escherichia coli. 1. Composition. *Biotechnol Prog* **9:** 539-47.

Wang H., Griffiths M.N., Burton D.R. and Ghazal P. 2000. Rapid antibody responses by low-dose, single-step, dendritic cell-targeted immunization. *Proc Natl Acad Sci U S A* **97:** 847-52.

Ward J.J., McGuffin L.J., Bryson K., Buxton B.F. and Jones D.T. 2004. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20:** 2138-9.

Wu D.Y., Ugozzoli L., Pal B.K., Qian J. and Wallace R.B. 1991. The effect of temperature and oligonucleotide primer length on the specificity and efficiency of amplification by the polymerase chain reaction. *DNA Cell Biol* **10:** 233-8.

Zhang L., Falla T., Wu M., Fidai S., Burian J., Kay W. and Hancock R.E. 1998. Determinants of recombinant production of antimicrobial cationic peptides and creation of peptide variants in bacteria. *Biochem Biophys Res Commun* **247:** 674-80.

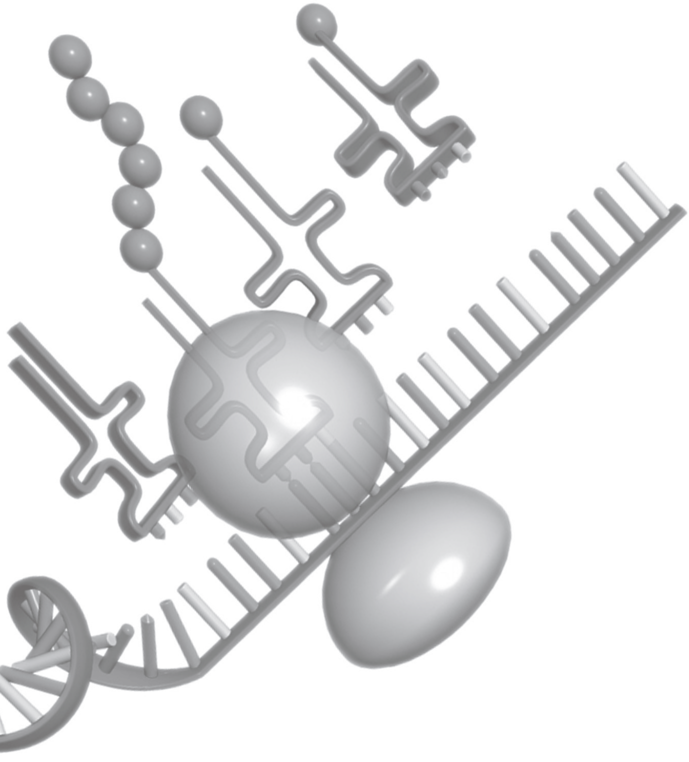
Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**: 3406-15.

5

Regionalized GC content of template DNA as a predictor of PCR success

Yair Benita[1], Ronald S Oosting[1], Martin Lok[1], Michael J Wise[2] and Ian Humphery-Smith[1]

Nucleic Acids Res 2003; 31 (16): e99



[1]Department of Pharmaceutical Proteomics, Utrecht Institute of Pharmaceutical Sciences, Utrecht University, Sorbonnelaan16, Utrecht, The Netherlands.

[2]Department of Genetics, Cambridge University, Cambridge CB2 3EH, England.

ABSTRACT

A set of 1,438 human exons was subjected to nested PCR. The initial success rate using a standard PCR protocol required for ligation-independent cloning was 83.4%. Logistic regression analysis was conducted on 27 primer- and template- related characteristics, of which most could be ignored apart from those related to the GC content of the template. Overall GC content of the template was a good predictor for PCR success, however, specificity and sensitivity values for predicted outcome were improved to 84.3% and 94.8% respectively when regionalized GC content was employed. This represented a significant improvement in predictability with respect to GC content alone ($P \leq 0.001$; Chi square) and is expected to increase in relative sensitivity as template size increases. Regionalized GC was calculated with respect to a threshold of 61% GC content and a sliding window of 21 base pairs across the target sequence. Fine-tuning of PCR conditions is not practicable for all target sequences whenever a large number of genes of different lengths and GC content are to be amplified in parallel, particularly if total ORF or domain coverage is essential for recombinant protein synthesis. Thus, the present method is proposed as a means of grouping subsets of genes possessing potentially difficult target sequences so that PCR conditions can be optimized separately in order to obtain improved outcomes.

INTRODUCTION

The recent mapping of the human, mouse, fly, yeast and other genomes has paved the way to an era of massive intra- and inter-genomic comparisons. In parallel, biomedical research laboratories, biotechnology and pharmaceutical companies have developed high-throughput methods for genomics and proteomics applications (Chapman 2003). Many of these methods depend upon amplification of nucleic acids by PCR (Markoulatos et al. 2002; Myakishev et al. 2001; Vieux et al. 2002; Zhang et al. 2001).

PCR requires a DNA template and a pair of primers flanking the target DNA. An important parameter to be considered when selecting PCR primers is the ability of the primers to form a stable duplex exclusively with the specific site on the target DNA. The use of the nearest neighbor thermodynamic parameters for computing DNA or RNA duplex stability has been shown to produce reliable predictions (Breslauer et al. 1986; Freier et al. 1986; Sugimoto et al. 1996; Wu et al. 1991). These methods calculate the melting temperature (Tm) of the primers, which is correlated with the GC/AT ratio of the primers. Typically, primers should have a GC/AT ratio similar to or higher than that of the amplified template (Rychlik 1993). Other considerations that increase the specificity of PCR include: (i) avoidance of complementarity at the 3′ termini of the primers, as this promotes the formation of primer-dimer artifacts; and (ii) avoidance of stable self-complementary hairpin loops that increase primer stability (Rychlik 1993).

The DNA template used for PCR is often overlooked when compared with the effort put into primer design. The most commonly used parameters that relate to the DNA template are the PCR product size and melting temperature of the product (Rozen and Skaletsky 2000; Rychlik 1993; Rychlik et al. 1990). However, it is known that DNA templates with very high or very low GC/AT ratio can be difficult to amplify (Baskaran et al. 1996; Chenchik et al. 1996; Varadaraj and Skinner 1994).

PCR has become a well-understood *in vitro* process (Mullis et al. 1994). Many tools exist that help to achieve a high yield of PCR products, such as, primer design software (Rozen and Skaletsky 2000; Rychlik and Rhoads 1989), optimization kits and well-characterized protocols (McPherson and Møller 2000; White 1993). However, these tools are often designed for a small number of reactions, or indeed a specific gene whereby the temperature and/or ion concentrations are varied to achieve maximal recovery of desired product (McPherson and Møller 2000). This is not feasible when hundreds of genes are to be amplified in parallel.

Several recent studies have evaluated the success of primer extension for genotyping (Vieux et al. 2002; Yuryev et al. 2002) and for generation of gene sequence tags (Varotto et al. 2001). Vieux et al. (Vieux et al. 2002) reported a 96% success rate in PCR using a very strict primer selection strategy combined with stringent PCR conditions for analysis of single nucleotide polymorphisms. Theses applications have the luxury of scanning long nucleotide sequences until the optimal primers are found. However, amplifying a particular DNA sequence of interest does not usually allow a stringent primer selection strategy, especially if the target sequence is a few hundred base-pairs in length or contains the whole or specific portions of open reading frames (ORF's) for recombinant protein synthesis (Albala et al. 2000; Braun et al. 2002; Christendat et al. 2000; Hammarstrom et al. 2002). The latter are thought to become increasingly important in a proteomics context.

Here we report on the amplification of 1438 human exons and efforts to establish a suitable predictor of PCR outcome.

MATERIALS AND METHODS

Selection of exons

We randomly selected 1438 human ORF's from disease-related genes available in publicly-accessible clone libraries in late 2001 and retrieved their DNA coding sequence from GenBank (http://www.ncbi.nlm.nih.gov). Coding sequences were compared to the human genome (GenBank build 25) using BLAST (Altschul et al. 1997) and the exons were extracted and set in frame. For ORF's containing multiple exons, the first was discarded to reduce the likelihood of a signal protein and from the remaining exons the longest was chosen.

Primer design strategy

We selected by default the first and last 21 nucleotides of each target sequence as the primers and modified each primer only if more than 4 G's or 4 C's were present in the last five nucleotides of the 3′ end, or if more than 3 consecutive T's were present at the 3′ end. In such cases, up to five nucleotides were removed from the 3′ end, allowing a minimum primer length of 16 nucleotides. This study was conducted with a view to subsequent cloning in the Gateway ™ system (Invitrogen). Therefore, two long adapters, named attB1 and attB2, had to be attached to both sides of the PCR product in a two-step procedure. Firstly, an oligonucleotide of 14 bases was attached to the 5′ end of the forward primer (AAAAAGCAGGCTTG) and an oligonucleotide of 13 bases was attached to the 5′ end of the reverse primer (AGAAAGCTGGGTA). Secondly, two universal primers were employed that bound to the adapters from the first PCR. The forward universal primer

GGGGACAAGTTTGTACAAAAAAGCAGGCTTG and the reverse universal primer GGG GACCACTTTGTACAAGAAAGCTGGGTA were used to complete the attB1 site and attB2 site. All primers were synthesized by Sigma Genosys.

Polymerase Chain Reaction

Genomic DNA was isolated from purified human white blood cells using a Genomic tip™ 500/g from Qiagen. A two-step PCR was performed in 96 well plates with a GeneAmp PCR system 9700 from Applied Biosystems. The standard PCR conditions were: 0.1 µg of template DNA, 0.05 µL TaKaRa Ex Taq, 1 µl 10X Ex Taq Buffer (2 mM Mg^{++}), 0.8 µl dNTP mixture (2.5 mM each) and 0.5 µM of each primer in a 10 µl reaction mixture. In all PCR cycles, denaturation lasted 30 seconds at 94°C and polymerization 2 minutes at 72°C. The annealing step was for 30 seconds at varying temperatures, namely, at 58°C in the first PCR and at 45°C for 5 cycles followed by 65°C for 25 cycles in the second PCR. PCR products were visualized with 0.5 µL/ml ethidium bromide on a 1.2% agarose gel. Images were taken using GeneGenious from Syngene® and analyzed with the bundled GeneTools software.

Logistic Regression

A stepwise backward likelihood ratio (LR) logistic regression was performed with SPSS version 10. Entry and removal p-values were set to ≤ 0.05. The receiver operating characteristic (ROC) curve was used as a measure of model performance. It was employed graphically to represent the trade-off between false positive and false negative rates for every possible cut off. The false positive rate was plotted on the x-axis and the true positive rate (1 - the false negative rate) on the y-axis. The area under the curve was of primary interest as it measured the correlation between the category predicted by the test and the true category into which the case falls (Beck and Shultz 1986; Centor and Schwartz 1985).

Informatics

Software for sequence analysis of primers and DNA template was written in Python (http://www.python.org) and all data and results were stored in a FileMaker database (http://www.filemaker.com). SPSS software version 10 was used for data analysis and statistical modeling. The parameters employed for the study of primers and DNA template are summarized in Table 1. In all statistical tests the primers were labeled 1 and 2 according to their GC content. Primer 1 is the primer with higher GC content of the two primers and not necessarily the forward primer.

Regionalized GC content within template DNA was calculated using a sliding window of 21 nucleotides, shifted one nucleotide at a time. The results were plotted and the area under the GC curve (AUC_{GC}) above a 61% threshold was calculated using the trapezoid method (Figure 1). A high GC content region was considered significant if it was above 61% for ten consecutive windows. Similarly, regionalized Tm and the area under

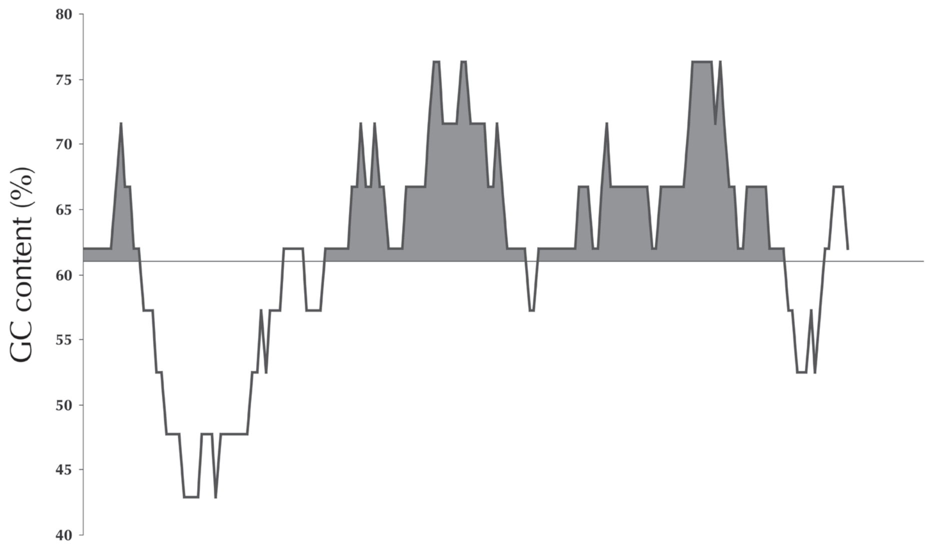


Figure 1: A sliding window of GC content plotted against the DNA sequence of an example template. The gray regions represent areas above a 61% threshold for GC content and measured across 31 base pairs, i.e. x10 sliding windows of 21 base pairs.

Table 1: Description of parameters used for analyzing PCR primers and template

Parameter Name	Description	Reference
T_m^{primer}	Melting temperature of the primers	Freier et al. 1986; Rychlik 1993
IntStab3	Internal stability of the primer at the 3′ end	Freier et al. 1986; Rychlik 1993
IntStab5	Internal stability of the primer at the 5′ end	
IntStabD	IntStab5 – IntStab3	
GC_{primer}	Primer (G+C) / Primer length	
GC_DIFF	GC_{primer} / $GC_{Template}$	
SigHits	Number of significant hits when comparing the primer sequence with the human genome using BLAST. A blast-hit was considered significant when 10 identical nucleotides occurred at the 3′ end.	
Bend	Bending value at the 3′ end of the primer	EMBOSS, banana Rice et al. 2000
Curve	Curvature value at the 3′ end of the primer	
SelfAny	The maximum local alignment score when testing a primer for annealing with itself or with the other primer. Computed by Primer3.	Rozen and Skaletsky 2000
SelfEnd	The maximum 3′-anchored global alignment score when testing a primer for annealing with itself or with the other primer. Computed by Primer3.	
$T_m^{Product}$	Melting temperature of the PCR template	Baldino et al. 1989
$GC_{template}$	Template (G+C) / Template length	
T_mDiff	$\|T_m^{primer1} – T_m^{primer2}\|$	
T_a^{OPT}	Optimal temperature for PCR	Rychlik and Rhoads 1989
Dimer	Highest possible duplex stability between both primers, calculation based on free energy values.	Freier et al. 1986
MaxCurve	Highest DNA curvature in the PCR template.	EMBOSS, banana (Rice et al. 2000)
AUC_{Tm}	Area under the T_m curve and above 75°C of the PCR template	
AUC_{GC}	Area under the GC curve and above 65% of the PCR template	
$ratio_{GC}$	Number of GC windows with values above 65% divided by the length of the PCR template	
$ratio_{Tm}$	Number of T_m windows with values above 75°C divided by the length of the PCR template	
$NormAUC_{GC}$	$ratio_{GC}$ x AUC_{GC}	
$NormAUC_{Tm}$	$ratio_{Tm}$ x AUC_{Tm}	
MinDist	Shortest distance from either ends of the PCR template and the first high GC region	
MIN_{GC}	Low value of the GC content of the first and last 60 nucleotides of the PCR template.	
MAX_{GC}	High value of the GC content of the first and last 60 nucleotides of the PCR template.	
SIZE	PCR product length	

the Tm curve (AUC$_{Tm}$) above a threshold of 74°C were calculated. The thresholds for both the GC curve (AUC$_{GC}$) and the Tm curve (AUC$_{Tm}$) were chosen initially as 65% and 75°C so as to reflect population extremes. Subsequently, these threshold values were made more precise with respect to their ability to discriminate between 'good' and 'failed' groups for all integer values between 50% -70% and 65°C - 85°C respectively, while employing the LR logistic regression. Table 1 summarizes the methods and parameters employed for statistical analysis, while associated software is available from: http://wwwcmc.pharm.uu.nl/moret/pub/benita.

RESULTS

Primers and template properties

Out of 2,876 primers, 2,501 were not altered while one to four nucleotides were removed from the 3′ end of the remaining 375. Properties employed for primer evaluation are shown in Table 2. A wide range of values was allowed for each primer property, yet, average values of Tm, GC content and internal stability were well within the recommended range, as defined by Rychlik (Rychlik 1993; Rychlik 1995) and McPherson (McPherson and Møller 2000). However, when combined parameters are examined, for example, Tm and 3′ end internal stability together, results are less clear-cut. In the latter example, only 60% of the primers were within the recommended range and only 37% of the primer pairs were both within the range. Primers had 4.8 significant hits, on average, when compared to the human genome using BLAST. A search of the entire human genome for potential PCR products that required the primers to be on opposite strands and not more than 2000 nucleotides apart, predicted that only one PCR product could be formed by *in*

Table 2: Properties of the 2876 primers employed

	Tm (°C)	GC content (%)	GC Diff	IntStab3 (Kcal/mol)	IntStab5 (Kcal/mol)	IntStabD	SelfAny	SelfEnd
Average	60.6±7.12	49.0±11.3	0.95±0.18	7.6±1.2	7.83±1.35	0.23±1.73	5.1±1.9	2.6±2.2
Range	36.1 - 86.8	14.0 – 86.0	0.34-1.73	5-13.1	4.8 - 13	-5.17 – 6.3	0 - 14	0 - 12
Recommended	55-70	30-70	> 1	< 9	N.A.	> 0	< 8	< 3

silico predictions. *In silico*, all pairs of primers generated a single target PCR product. A combined analysis of both primers and the PCR template was performed to evaluate the success of the reaction as shown in Table 3. The observed range was very wide for most of the parameters. Analysis of DNA curvature was included to identify DNA structural oddities that might have affected the ability of *Taq polymerase* to duplicate the template.

Table 3: Template and primer properties

	Ta Opt (°C)	Tm Diff (°C)	Lowest GC Diff	ΔG dimer (Kcal/mol)	Tm product (°C)	GC content (%)	DNA max Curvature (deg)
Average	56.5±4.8	7.3±5.7	0.8±0.2	4.5±7.0	78.2±3.9	51.9	32.4
Range	32.8 – 69.6	0.01 – 32.8	0 – 1.4	0 –30.5	68.2 – 89.6	31.2 – 77.1	9.7 –120.4
Recommended	58	≤ 5	≥ 1	≤ 12	None	30-70	None

PCR

A PCR product with an expected size higher than 350 bp was considered ‘good’ if the observed band was 10% longer or shorter than the expected size. A maximal deviation of 15% was allowed for smaller products, due to the inherent insensitivity of on-gel mobility measurements. All bands below 120 base pairs were discarded and interpreted as representing primer dimmers or PCR artifacts. A band of the expected size was observed for 1,226 (83.4%) sequences. The other 212 (14.7%) failed in duplicate experiments to produce the correct band size, 69 (32.5%) had no product at all and 44 (20.7%) were associated with a product of incorrect size. Of all the ‘good’ products, 858 (70%) had one clear band and the other, 305 (25%), 54 (4.4%) and 8 (0.75%) had two, three and four bands, respectively.

Numerical analysis

Two data sets were created for the analysis, data set A contained 212 sequences that failed to PCR twice and data set B contained 318 sequences, which produced twice a clear visible band. We avoided including too many samples in data set B since it could bias the statistical analysis. Seventy percent of the data in each set was used for statistical analysis as selected by a random function, while the remainder was used as a test set for the prediction model. The mean values of groups A and B for the parameters described in Table 1 were compared using a one-way ANOVA. The ANOVA results showed that Tm and GC content are the most significant parameters at the primer level and parameters that are correlated with total GC content of the template are the most significant at the template level (Figure 2). All the primer and template parameters were used in a stepwise backward likelihood ratio logarithmic regression. Although total GC content, GC ratio, $T_m^{Product}$ and T_a^{opt} are the most significant parameters in the ANOVA test, $NormAUC_{GC}$ and $NormAUC_{Tm}$ were shown to be much better predictors in the logistic regression. A logistic regression for each variable was performed separately and the goodness of fit was assessed by –2 log likelihood. The strongest single predictor of success and failure of PCR, using the logistic regression model was $NormAUC_{GC}$ (Figure 3). The best logistic regression model contained both the primer with lower GC content ($GC_{primer2}$) and the $NormAUC_{GC}$. Wald $\chi 2$ values for $GC_{primer2}$, $NormAUC_{GC}$ and the constant were respectively 19.2, 40.1, and 38.2, each with similar degrees of freedom. Thus, a good level of confidence was obtained for the expected PCR success, as shown by the following equation:

[redacted]

where p is the probability of a successful PCR and $GC_{primer2}$ and $NormAUC_{GC}$ are the parameters described in Table 1. The area under the ROC curve was 0.87, and the Nagelkerke R square was 0.58. Both are high and suggest the model's performance is good. A PCR is predicted successful for $p \geq 0.5$. Using this equation on our test set (n = 160; i.e. 30% of data set A + B), 86.3% PCR were predicted correctly. Sensitivity of the model is the probability of correctly predicting a positive example, while the specificity is the probability that a positive example is correct (Baldi and Brunak 2001). The specificity and sensitivity values of the test set were 84.3% and 94.8% respectively; and 94.9% and 85.2% respectively for the 1,438 PCR reactions examined in this study.

The logistic regression equation was able to predict that for a given value of $NormAUC_{GC}$ a reduction of the GC content of the primer should increase the probability of PCR success. This was due to the significantly lower $GC_{primer2}$ of group B when compared to that of group A, at $NormAUC_{GC} \leq 340$. The mean values of $GC_{primer2}$ for $NormAUC_{GC} \leq$

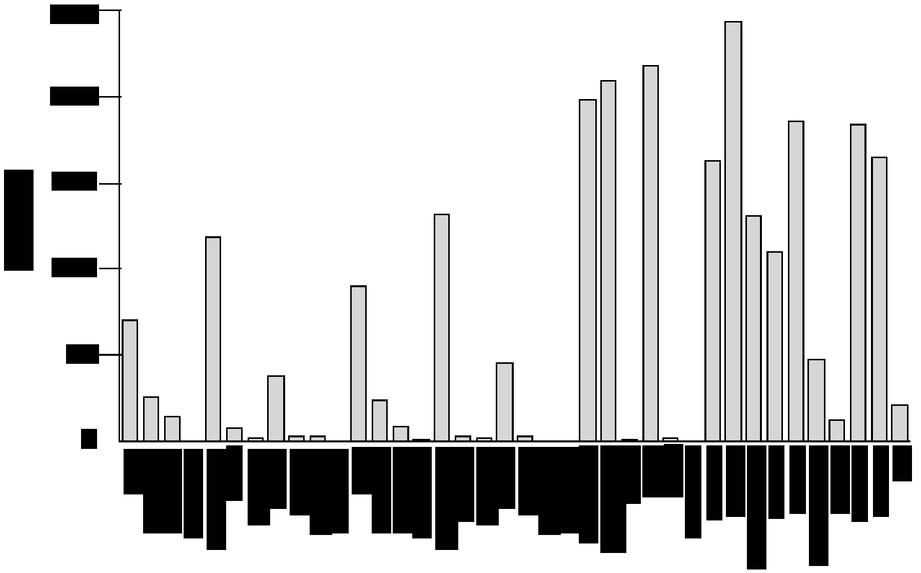


Figure 2: F-values for a one-way ANOVA computed from a learning set of 370 DNA target templates. Large F-values represent a significant difference between the failed and successful PCR. p=0.05 for F-value=3.87 and p=0.001 for F-value=11.

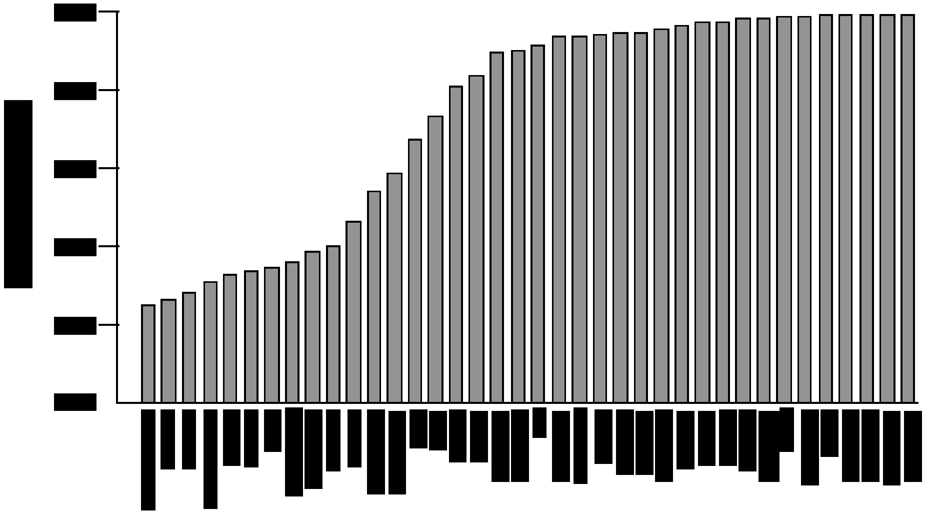


Figure 3: Likelihood values -2 log obtained by logistic regression for each variable separately. These values are commonly used to indicate the goodness of fit. The lower the value the better the model.

340 for groups A and B were 48%±10 and 44%±9 ($p < 0.05$), respectively, corresponding to a 2°C difference in mean $T_m^{primer2}$ values. Nevertheless, for $NormAUC_{GC} > 340$ the probability of PCR success was significantly reduced even for low $GC_{primer2}$ values, since 67.2% of the sequences in group A possessed a $NormAUC_{GC} > 340$ compared with only 5.2% for group B. Therefore, PCR success was also predicted based on $NormAUC_{GC}$ alone with an upper threshold of 340. The complete set of 1438 sequences was divided into three groups, sequences with $NormAUC_{GC} \leq 340$; $NormAUC_{GC}$ between 340 and 750 (95 percentile of successful PCR); and $NormAUC_{GC} > 750$. In the first category ($n = 1143$), the success rate was 93.8%, while in the second ($n = 139$) and third ($n = 156$) it was 71.2% and 35.2% respectively. Thus, the high predictive value of the index, $NormAUC_{GC} \leq 340$, was clearly demonstrated.

DISCUSSION

For more than a decade primer design has evolved into an efficient and mature science. Although the primer sequence can be modified by biologists, the target DNA cannot. Therefore, little attention is usually afforded to the analysis of the PCR template prior to experimental procedures, i.e. apart from its relevance to primer design. Faced with the challenge of large-scale PCR, protocol optimization becomes increasingly important, yet is increasingly problematic due to the variation in template sequence and length. As a result, overall PCR success rates can be compromised. In this study, we found regionalized GC content to be a good predictor of PCR success across multiple templates. Indeed, any parameter able to be correlated with the GC content of the PCR template, such as $T_m^{Product}$ and T_a^{Opt}, was statistically significant when PCR success and failure were compared. However, normalized area under the template GC curve ($NormAUC_{GC}$) was seen to be a better predictor for PCR success ($P \leq 0.001$; Chi square) for the total data set of 1438 PCR's. $NormAUC_{GC}$ was much more sensitive to fluctuations in GC content than other methods that simply relied on averaging overall GC. The performance of this predictor is expected to improve as template size increases due to the greater likelihood for problematic regions to occur within a given template.

The evidence presented here would suggest that the primer was most often not the cause of PCR failure, but rather the template, i.e. because all primers met similar stringency demands. In all cases, the average values of 63% and 52% for $GC_{primer1}$ and $GC_{primer2}$ of failed PCR reactions were acceptable, even if the most stringent primer design criteria were employed (Vieux et al. 2002). Furthermore, Rychlik et al. (Rychlik et al. 1990) showed that primer design was significant for a low number of PCR cycles, while this diminished after 25 cycles. When employing nested primer PCR of 30 cycles for each reaction, as here for ligation-independent cloning experiments, strong amplification can be expected to depend less upon stringent primer design due to the addition of 14 and 13 bases to the 5′ ends of the specific forward and reverse primers, respectively, and the associated increase in affinity of primers for template. Thus, provided obvious homologies and self-annealing attributes of primers are minimized, then most 20-mer strings will be associated with sufficient target specificity. As a consequence, our results would suggest that more effort should be put towards analysis of the PCR template. Stringent primer design might result in high amounts of very pure PCR product, but it comes at the expense of sequence coverage. The latter is most important when entire ORF's or domains are being targeted and template coverage is essential for recombinant protein synthesis.

For templates possessing a $NormAUC_{GC} > 340$, it is predicted that the success of PCR will be more dependant on a suitable protocol than on primer selection of primers. Therefore, when faced with the task of large-scale PCR, we recommend dividing the samples to three groups and subsequently optimizing the PCR for successful outcomes in each of the following categories: sequences with $NormAUC_{GC} \leq 340$; $NormAUC_{GC}$ between 340 and 750; and $NormAUC_{GC} \geq 750$. The first group is likely be successfully amplified using standard PCR protocols, and as a result primer stringency may be relaxed without deleterious effects, thereby allowing maximal target sequence coverage. The second and third groups should each be optimized in turn with increasing attention being given to protocol and primer design.

In summary, the $NormAUC_{GC}$ of a PCR template was found to represent a more sensitive predictor of PCR outcome than parameters previously described, while its predictive value as an improvement on GC content alone is likely to increase concomitantly with template size. Although the learning set examined during this study was derived from nested primer PCR, the index, NormAUC$_{GC}$, is expected to maintain its relevance for standard PCR experiments, as most primer-related failures probably occur during initial cycles only.

ACKNOWLEDGEMENTS

We wish to thank Glaucus Proteomics B.V. for financial and logistic support for this study, while Rogier Donders, Ed Moret and David Gestel are acknowledged for the valuable input.

REFERENCES

Albala, J.S., K. Franke, I.R. McConnell, K.L. Pak, P.A. Folta, B. Rubinfeld, A.H. Davies, G.G. Lennon, and R. Clark. 2000. From genes to proteins: high-throughput expression and purification of the human proteome. *J. Cell. Biochem*. **80**: 187-191.

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389-3402.

Baldi, P. and S. Brunak. 2001. Bioinformatics: The machine learning approach. MIT Press, London.

Baldino, F., Jr., M.F. Chesselet, and M.E. Lewis. 1989. High-resolution in situ hybridization histochemistry. *Meth. Enzymol*. **168**: 761-777.

Baskaran, N., R.P. Kandpal, A.K. Bhargava, M.W. Glynn, A. Bale, and S.M. Weissman. 1996. Uniform amplification of a mixture of deoxyribonucleic acids with varying GC content. *Genome Res* **6**: 633-638.

Beck, J.R. and E.K. Shultz. 1986. The use of relative operating characteristic (ROC) curves in test performance evaluation. *Arch. Pathol. Lab. Med.* **110**: 13-20.

Braun, P., Y. Hu, B. Shen, A. Halleck, M. Koundinya, E. Harlow, and J. LaBaer. 2002. Proteome-scale purification of human proteins from bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **99**: 2654-2659.

Breslauer, K.J., R. Frank, H. Blocker, and L.A. Marky. 1986. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. U.S.A.* **83**: 3746-3750.

Centor, R.M. and J.S. Schwartz. 1985. An evaluation of methods for estimating the area under the receiver operating characteristic (ROC) curve. *Med Decis Making* **5**: 149-156.

Chapman, T. 2003. Lab automation and robotics: Automation on the move. *Nature* **421**: 661-666.

Chenchik, A., L. Diachenko, F. Moqadam, V. Tarabykin, S. Lukyanov, and P.D. Siebert. 1996. Full-length cDNA cloning and determination of mRNA 5′ and 3′ ends by amplification of adaptor-ligated cDNA. *Biotechniques* **21**: 526-534.

Christendat, D., A. Yee, A. Dharamsi, Y. Kluger, M. Gerstein, C.H. Arrowsmith, and A.M. Edwards. 2000. Structural proteomics: prospects for high throughput sample preparation. *Prog. Biophys. Mol. Biol.* **73**: 339-345.

Freier, S.M., R. Kierzek, J.A. Jaeger, N. Sugimoto, M.H. Caruthers, T. Neilson, and D.H. Turner. 1986. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. U.S.A.* **83**: 9373-9377.

Hammarstrom, M., N. Hellgren, S. van Den Berg, H. Berglund, and T. Hard. 2002. Rapid screening for improved solubility of small human proteins produced as fusion proteins in *Escherichia coli*. *Protein Sci.* **11**: 313-321.

Markoulatos, P., N. Siafakas, and M. Moncany. 2002. Multiplex polymerase chain reaction: a practical approach. *J. Clin. Lab. Anal.* **16**: 47-51.

McPherson, M.J. and S.G. Møller. 2000. PCR. BIOS Scientific Publishers Ltd, Oxford.

Mullis, K.B., F. Ferre, and R.A. Gibbs. 1994. The polymerase chain reaction. Birkhauser, Boston.

Myakishev, M.V., Y. Khripin, S. Hu, and D.H. Hamer. 2001. High-throughput SNP genotyping by allele-specific PCR with universal energy-transfer-labeled primers. *Genome Res*. **11**: 163-169.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16: 276-277.

Rozen, S. and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol*. **132**: 365-386.

Rychlik, W. 1993. Selection of primers for polymerase chain reaction. In PCR protocols: current methods and applications (ed. B.A. White), pp. 31-40. Humana Press, Totowa.

Rychlik, W. 1995. Selection of primers for polymerase chain reaction. *Mol. Biotechnol.* **3**: 129-134.

Rychlik, W. and R.E. Rhoads. 1989. A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. *Nucleic Acids Res.* **17**: 8543-8551.

Rychlik, W., W.J. Spencer, and R.E. Rhoads. 1990. Optimization of the annealing temperature for DNA amplification in vitro. *Nucleic Acids Res.* **18**: 6409-6412.

Sugimoto, N., S. Nakano, M. Yoneyama, and K. Honda. 1996. Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res*. **24**: 4501-4505.

Varadaraj, K. and D.M. Skinner. 1994. Denaturants or cosolvents improve the specificity of PCR amplification of a G + C-rich DNA using genetically engineered DNA polymerases. *Gene* **140**: 1-5.

Varotto, C., E. Richly, F. Salamini, and D. Leister. 2001. GST-PRIME: a genome-wide primer design software for the generation of gene sequence tags. *Nucleic Acids Res.* **29**: 4373-4377.

Vieux, E.F., P.Y. Kwok, and R.D. Miller. 2002. Primer design for PCR and sequencing in high-throughput analysis of SNPs. *Biotechniques* **Suppl**: 28-30, 32.

White, B.A. 1993. PCR protocols: current methods and applications. Humana Press, Totowa.

Wu, D.Y., L. Ugozzoli, B.K. Pal, J. Qian, and R.B. Wallace. 1991. The effect of temperature and oligonucleotide primer length on the specificity and efficiency of amplification by the polymerase chain reaction. *DNA Cell Biol.* **10**: 233-238.

Yuryev, A., J. Huang, M. Pohl, R. Patch, F. Watson, P. Bell, M. Donaldson, M.S. Phillips, and M.T. Boyce-Jacino. 2002. Predicting the success of primer extension genotyping assays using statistical modeling. *Nucleic Acids Res.* **30**: E131-131.

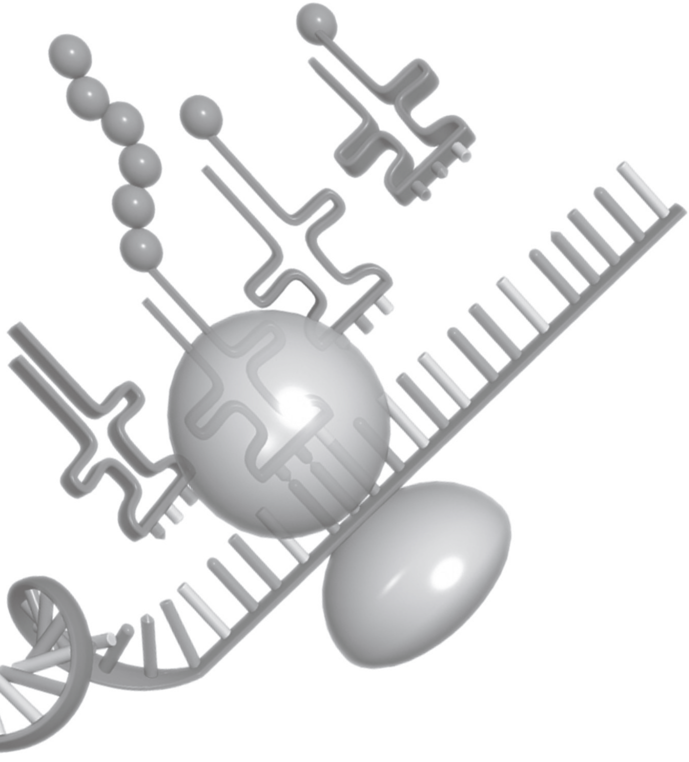
Zhang, Y., Y. He, and E.S. Yeung. 2001. High-throughput polymerase chain reaction analysis of clinical samples by capillary electrophoresis with UV detection. *Electrophoresis* **22**: 2296-2302.

6

Analysis of High-Throughput Protein Expression in *Escherichia Coli*

Yair Benita[1], Michael J Wise[2], Martin Lok[3], Ian Humphery-Smith[4] and Ronald S Oosting[1]

Submitted for publication



[1]Department of Psychopharmacology, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, The Netherlands

[2]The University of Western Australia, Crawley, Australia

[3]Department of Pharmaceutics, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, The Netherlands

[4]Biosystems Informatics Institute, Newcastle, United Kingdom

ABSTRACT

The ability to efficiently produce hundreds of proteins in parallel is the most basic requirement of many aspects of proteomics. Overcoming the technical and financial barriers associated with high-throughput protein production is essential for the development of an experimental platform to query and browse the protein content of the cell (e.g. protein and antibody arrays). Proteins are inherently different one from another in their physicochemical properties, therefore, no single protocol can be expected to successfully express most of the proteins. Instead of optimizing a protocol to express a specific protein, we employed sequence analysis tools to estimate the probability of a specific protein to be expressed successfully using a given protocol, thereby, avoiding *a priori* proteins with a low success probability. A set of 547 proteins, to be used for antibody production and selection, was expressed in *Escherichia coli* using a high-throughput protein production pipeline. Protein properties derived from sequence alone were correlated to successful expression and general guidelines are given to increase the efficiency of similar pipelines. A second set of 68 proteins was expressed to investigate the link between successful protein expression and inclusion body formation. More proteins were expressed in inclusion bodies, however, the formation of inclusion bodies was not a requirement for successful expression.

INTRODUCTION

The completion of the human genome project and the biotechnical advances in the genomics field have radically transformed biological and medical research. We now have the ability to monitor the mRNA expression of thousands of genes simultaneously in cells and tissues. However, it is the proteins encoded by these genes that carry out most biological functions. The proteome is much more daunting in size and complexity than the genome and to understand how cells work we must examine which proteins are present, how they interact with each other and what they do. The study of proteins is difficult because they are each distinctively different from the other and are usually present in tissue in very low amounts. In the absence of a PCR equivalent, it has been suggested to call upon affinity ligands, such as monoclonal antibodies, for detection and identification of proteins (Humphery-Smith 2004). Regardless of the specific affinity ligand used, purified proteins must first be acquired in large quantities for generation and/or selection of specific affinity ligands. Thus, there is a need to define expression and purification conditions that are amenable to hundreds or even thousands of proteins in parallel. However, owing to the significant differnce in physicochemical properties, the success rate of high-throughput protein production is often too low, resulting in an increased financial burden and technical constraints on such projects.

Several groups have previously attempted high-throughput expression of proteins or protein fragments. High-throughput is defined as the ability to automate protein production, commonly using a 96-well format. Braun et al. expressed 336 randomly selected human cDNAs in *E. coli* and purified successfully 60% under denaturing conditions using a His_6 construct and 50% under non-denaturing conditions using a GST construct (Braun et al. 2002). Luan et al. expressed 10,176 *Caenorhabditis elegans* proteins using a robotic pipeline and observed an overall expression of 50% (15% in soluble form) (Luan et al. 2004). Agaton et al. reported a success rate of 76% for the expression of 142 human proteins in *E. coli* (Agaton et al. 2003). Other groups reported success rates in the range of 60%-80% (Christendat et al. 2000; Pizza et al. 2000; Dobrovetsky et al. 2005).

The three dimensional structure of a protein can often provide functional clues, primarily by detecting structural homology with a protein of known function (Cort et al. 1999; Zarembinski et al. 1998). Structural proteomics attempts to determine protein structure on a genome-wide scale. It not only requires high-throughput expression of target proteins but also that the proteins be produced soluble, correctly folded and suitable for X-ray crystallography or NMR studies. Previous attempts to produce proteins on a large scale for structural studies resulted in success rates of ~10% (Bertone et al. 2001; Goh et al. 2003). This low success rate motivated studies that attempted to link a protein's primary sequence to its propensity to be soluble upon over-expression in *E. coli* (Bertone et al. 2001; Goh et al. 2003; Idicula-Thomas and Balaji 2005; Shimada et al. 2005). On the other hand, protein production for affinity ligands does not necessarily require the heterologous protein to be soluble. Agaton et al. reported a success rate of 56% for eliciting affinity-purified antibodies against proteins that were expressed in *E. coli* and purified under denaturing conditions (Agaton et al. 2003). In this respect protein production for affinity ligands is significantly less demanding than production for structural studies. To better cope with the financial constraints of high-throughput protein production, it would be beneficial to identify *a priori* proteins that are likely to fail expression in a pipeline designed for affinity ligands target generation. While prediction of protein solubility upon over-expression has drawn scientific attention, prediction of successful expression has been largely disregarded. Prediction of protein expression is bound to be more complicated, since expression can fail in any of several different steps from plasmid construct stability to the final purified protein. Many of those steps, such as mRNA decay, are not necessarily related to the primary protein sequence or to the physicochemical properties of the amino acids. Solubility, on the other hand, is more likely to be dependent on the amino acid composition of the protein.

In this study we present results on the expression of 547 recombinant proteins, produced as targets for affinity ligand generation, and investigate the link between their DNA and protein sequences and successful expression. Finally we investigate the relationship between solubility and expression level on a set of 68 human proteins.

METHODS

Selection of genes

We randomly selected 615 human ORFs - 547 for high-throughput expression and 68 for inclusion bodies analysis - from disease-related genes available in publicly accessible clone libraries in late 2001 and retrieved their DNA coding sequence from GenBank (http://www.ncbi.nlm.nih.gov). Coding sequences were compared with the human genome (Genbank build 25) using BLAST, and the exons were extracted and set in-frame. For ORFs containing multiple exon, the first was discarded to reduce the likelihood of a signal peptide, and from the remaining exons the longest was chosen. Primer selection criteria, genomic template and PCR protocols for our protein production pipeline were described previously (see chapter 5).

Protein expression and purification

The plasmid construct used, named HZS, contained a His_6 tag, a ZZ domain, a gateway compatible insert and a streptag. The ZZ domain is the tandem repeat dimer of the modified immunoglobulin binding domain of protein A of *Staphylococcus aureus* (Nilsson et al. 1987). The streptag (Skerra and Schmidt 2000) was constructed using custom oligos. Plasmid construction, gene cloning and bacterial transformation and induction have been described previously by our group (Zhao et al. 2005). As expression host, the *E. coli* BL21 codon-plus RP strain (Stratagene) was used. These cells contain extra copies of the argU and proL tRNA to enable expression of genes restricted by either AGG/AGA or CCC codons.

Protein purification for the high-throughput protein pipeline was performed under denaturing conditions. The bacteria were grown in 24 deep well plates. Each well contained 5 ml LB medium supplemented with 50 µg/ml of Ampicilin and chloramphenicol. At the end of the 4 hr IPTG-induction period, bacterial plates were centrifuged at 3500 rpm for 15 min. The supernatants were removed and the bacterial pellets were resuspended each in 1 ml lysis buffer containing 8M urea (lysis buffer: 100 mM NaH_2PO_4, 20 mM Tris, 10% Glycerol, 0.1% Tween 20, pH 8.0; 20 mM β-mercapto-ethanol plus one tablet of Complete protease inhibitor (Roche)). The content of each well was sonicated two times for 15 sec with 10 sec in between. Then the plates were centrifuged for 20 min at 3500 rpm. The next steps in the protein purification protocol were done using the Biorobot 8000 (Qiagen). Aliquots of 800 µl of the supernatants were transferred to a 96 well filterplate (Qiagen) containing 200 µl of Ni-NTA superflow that was washed once with 500 µl lysis buffer containing 8M urea before applying the supernatant. Then vacuum of 900 mBar was applied for 3 min. The resin was successively washed with 4M, 2M, 1M, and 0M solutions of urea in lysis buffer. After each wash step vacuum was applied for 1.5 min at 900 mBar and the flowthrough was discarded. Finally 1 ml of elution buffer (50 mM NaH_2PO_4, 300 mM NaCl, 250 mM imidazole, pH=8.0) was added to each well. After 10 min, vacuum was applied for 2 min at 700 mBar and the eluate was collected in a deep 96 well block.

Protein purification for inclusion body analysis was performed separately for the soluble and insoluble fractions of the bacterial lysate. The bacterial pellet from 10 ml of induced bacteria was resuspended in 600 µL of B-per (Pierce) containing 1 tablet of Complete protease inhibitor (Roche) per 25 ml, vortexed for 1 min at 3,000 rpm and centrifuged for 10 minutes in a standard tabletop microcentrifuge at 13,000 rpm and 4°C. The supernatant was removed and placed on a custom made column containing 100 µL Ni-NTA superflow. Columns were washed twice with 500 µL wash buffer (wash buffer: 50 mM NaH_2PO_4, 300 nM NaCl and 20 mM imidazole, pH=8.0) and eluted with 500 µL elution buffer. The remaining pellet of the lysed bacteria containing the insoluble fraction was resuspended by sonication (2x5 seconds) in 1 ml 8M urea and centrifuged for 10 minutes at 13,000 rpm. The supernatant was then placed on columns containing 100 µL Ni-NTA superflow and washed with 4M, 2M, 1M, and 0M solutions of urea in BR buffer. Proteins were eluted with 500 µL of elution buffer.

All proteins were visualized on Criterion Tris-HCl 12.5% precast poly-acrylamide gels from Bio-Rad and coomassie staining.

Sequence analysis

A sequence analysis module was written in Python (www.python.org) for this study and is being distributed as part of the SeqUtils module of biopython (www.biopython.org). An aromaticity score was calculated according to Lobry and Gautier (Lobry and Gautier 1994), and a protein instability index was calculated according to Guruprasad et al. (Guruprasad et al. 1990). Isoelectric point, charge and amino acid content (aliphatic, aromatic, polar, non-polar, charged, basic, acidic, small and tiny) were calculated using pepstats (Harrison 2000) from the Emboss package (http://emboss.sourceforge.net). Average and maximum protein flexibility were calculated according to Vinihen et al. (Vihinen et al. 1994). Protein disorder was calculated using FoldIndex (Prilusky et al. 2005) and from the output the longest disorder segment and the total number of residues in disorder segments were extracted. DNA sequence complexity was calculated using both nSEG (Wootton and Federhen 1996) and G1 (Wan and Wootton 2000). Protein secondary structure was assessed using garnier (Garnier et al. 1978) from the Emboss package and the fractions of helix, β-sheet, coil and turn were calculated from the output. Secondary structure of mRNA was predicted using mfold (Zuker 2003) and the most stable structure was selected (lowest ΔG). Protein low complexity was calculated using 0j.py (Wise 2001). Local GC content was calculated as previously described (see chapter 5 and also Figure 1). The grand average mean of hydrophobicity (GRAVY) was calculated according to Kyte and Doolittle (Kyte and Doolittle 1982). The GRAVY calculates an average value for the entire protein, which in many cases may be misleading. For example, on average a protein could be hydrophilic but still have a large internal hydrophobic region. Therefore, we calculated local hydrophilic and hydrophobic regions along the protein using the Kyte and Doolittle hydrophobicity plot generated with a sliding window of 11 amino acids. The areas under the curve (AUC) and above/below 0 were calculated using the trapezoid method, as shown in Figure 1 for GC content. The area above 0 was labeled "sum hydrophobic AUC" and the area below 0 was labeled "sum hydrophilic AUC". The single largest local hydrophobic or hydrophilic regions were located and labeled "max hydrophobic AUC" and "max hydrophilic AUC", respectively. The hydrophobic and hydrophilic AUC were

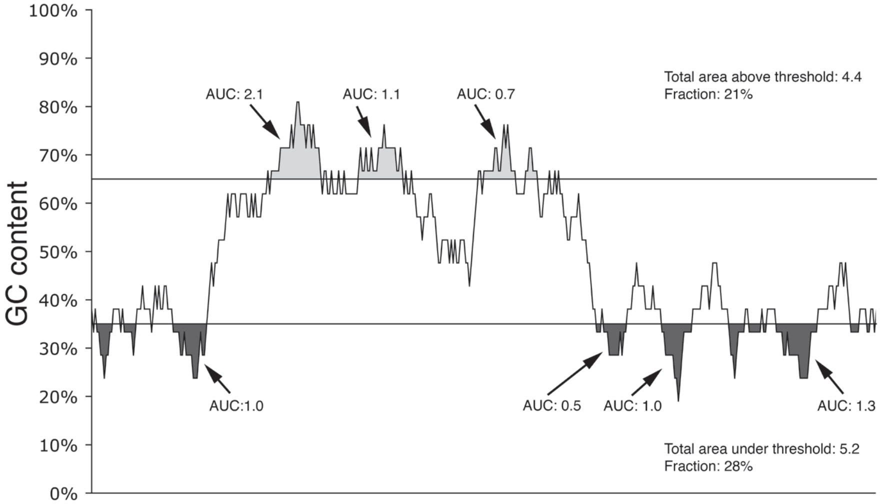


Figure 1: GC content of an exon that failed to be expressed. GC plot is generated using a sliding window approach of 21 bp. The area above 65% (light gray) and below 35% (dark gray) was calculated to locate high and low GC content regions. The fraction is the percentage of sequence that is above or below the threshold, i.e. 21% of the sequence has a GC content above 65%. The calculated areas of several large regions are shown.

also normalized by dividing the area by the total number of amino acids in the entire sequence. In addition, the hydrophobic to hydrophilic ratio was calculated, i.e. the ratio of the sum of all hydrophobic regions and the sum of all hydrophilic regions.

Codon usage was calculated according to Sharp and Li (Sharp and Li 1987). A set of 121 highly expressed *E. coli* proteins were selected from Swiss-2D PAGE (http://www.expasy.org/ch2d). All selected proteins were identified on a 2D protein gel and were present in large amounts with %vol average above 0.2 as calculated using the software Melanie (http://www.2d-gel-analysis.com). This set of proteins is available at http://wwwcmc.pharm.uu.nl/benita. The codon usage index was generated using a Python codon usage module (available through biopython) and the CAI for each gene was calculated. Regionalized CAI values were calculated using a CAI plot that was generated using a sliding window of 4 codons. The area below a threshold and above the curve was calculated for several thresholds (Figure 2). Both the sum of all areas and the single largest area were used for the analysis.

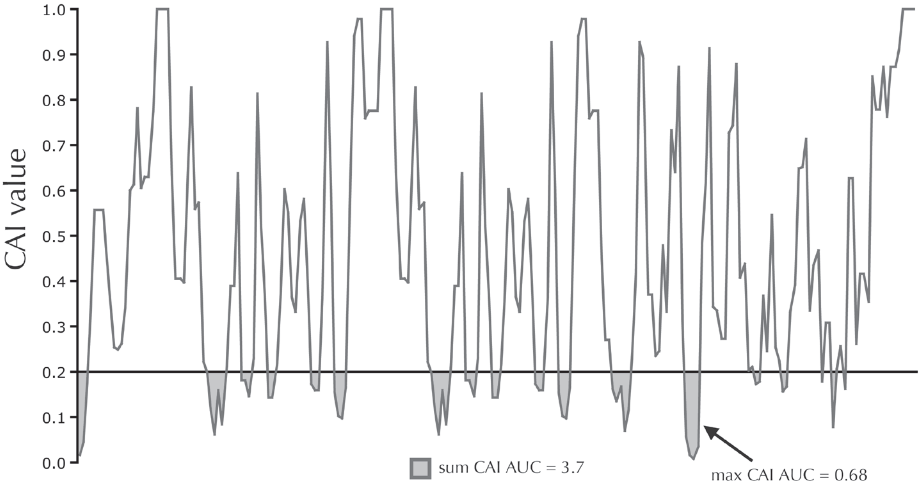


Figure 2: Plot of CAI value across a randomly selected human protein sequence, using a sliding window of 4 codons moving one codon at a time. Regionalized low CAI areas were calculated using the trapezoid method for regions below the threshold and above the CAI plot (light gray regions). Both the sum of all regions and the single largest region were used in the sequence analysis.

A modified version of the CAI, AAcai (amino acids-codon adaptation index), was introduced by taking into account amino acid shortage due to over-expression of a protein with an amino acid content different than the average *E. coli* protein. This attribute is based on the observation that ribosomes translating a heterologous mRNA may stall at positions calling for a tRNA that is largely deacylated because of the heavier than normal drain of its amino acid into protein (Kurland and Gallant 1996). In other words, the ribosome may stall even at an optimal codon if not enough amino acid is available to be loaded onto the tRNA. The average amino acid content of *E. coli* was calculated using the same set of 121 highly expressed *E. coli* proteins mentioned above. The amino acid content of each over-expressed sequence and the deviation from the average protein content were calculated. For each amino acid that was used more frequently than average the proportion of average usage to specific usage was calculated and the index used to calculate the CAI value was adjusted accordingly. For instance, if a specific protein had 20% alanine and the average *E. coli* protein had 10% alanine, the most abundant alanine codon was rescaled from 1 to 0.5 and all other alanine codons were adjusted accordingly. The probability of finding a loaded alanine tRNA was reduced by two fold due to the two fold increase in usage of alanine in the heterologous protein. Once the index values were calibrated to the amino acid usage, the exact same methods that were described above for CAI were employed.

Protein compositional bias was assessed using POPPs (Protein or Oligonucleotide Probability Profile) (Wise 2002), a suite of inter-related software tools which enable the user to discover statistically ‘unusual’ peptides. POPPs were created for each protein sequence versus the *E. coli* and human proteomes, scaled to a sequence length of 100 amino acids. The *E. coli* proteome was created using *E. coli* K-12 genome annotations (Genbank: NC_000913). The human proteome was fetched from the International Protein Index (Kersey et al. 2004). Redundant proteins with more than 99% and 98% similarity to another in the respective databases were removed using nrdb90 (Holm and Sander 1998).

T-tests and one-way ANOVA were performed using the stats.py and pstat.py modules (http://www.nmr.mgh.harvard.edu/Neural_Systems_Group/gary/python.html;also distributed as part of the SciPy package: http://www.scipy.org).

Decision trees are predictive models that are often used in machine learning. Here we used decision trees to classify proteins into expression groups based on DNA and protein sequence attributes. Inner nodes in the tree represented decision variables, e.g. degree of aromaticity, and leaf nodes represented the predicted expression groups. Decision trees were previously applied to similar data, linking sequence attributes to successful expression for structural analysis (Bertone et al. 2001; Goh et al. 2003). Decision trees were generated using the rpart module of the R statistical package (http://lib.stat.cmu.edu/R/CRAN). The decision trees were pruned using a complexity value of 0.05. To minimize the effect of unequal group sizes, a subset of randomly selected proteins were selected from the larger group equal in number to the small group. The process was repeated three times, generating three decision trees.

RESULTS

An initial set of 547 human exons, each representing a different gene, were transferred using Gateway high-throughput cloning system into the HZS vector construct. The protein fragments were a relatively small part of the entire recombinant protein. The average length of the protein insert was 76±29 amino acids, corresponding to 8.6±3.3 kDa. The constant part of the protein (HisZZ on the amino-terminal end and streptag on the carboxy-terminal end) was in total 170 aa long with a molecular weight of 19.45 kDa. The final HZS vectors containing the inserts were confirmed to be correct by observing the expected fragments on agarose gel after restriction-enzyme digestion. Protein expression was performed in *E. coli* BL21 and since the inserts were only protein fragments and not entire proteins, they were not expected to fold correctly. Therefore, all recombinant proteins were purified under denaturing condition. Protein expression was visualized using precast XT-gels and coomassie staining. The proteins were classified into one of five groups: (I) no visible bands; (II) faint bands with correct size; (III) faint bands with wrong size; (IV) strong bands with correct size; and (V) strong bands with wrong size (Figure 3). Classification into faint/strong was performed visually on a scanned gel image. Gray bands were labeled faint and black bands were labeled strong. In 77% of the proteins one or more bands were visible on the gel and overall in 58.5% of the proteins the expected size was observed. In all cases where a protein band was visible on the gel, the band was larger or equal to 17 kDa - the molecular weight of HisZZ, the constant amino-terminal of the recombinant protein.

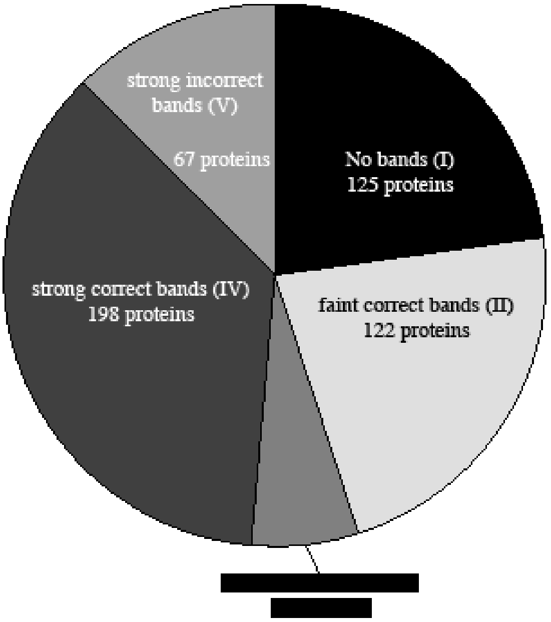


Figure 3: Summary of protein expression for 547 human proteins, as visualized on precast XT-gels (Bio-Rad). A band that was observed to be less than 10% smaller or larger than expected was labeled correct. Gray bands were labeled "faint" and black bands were labeled "strong".

Several DNA and protein sequence analysis methods were applied to the set of 547 human protein fragments. The objective of the analysis was to identify attributes that cause failure of expression in our protein production pipeline, thereby avoiding expression of proteins with a high failure probability. In the process of protein expression, a protein complex (mRNA polymerase or ribosome) scans an input molecule (DNA or mRNA) to produce an output molecule. Therefore, when applying several sequence analysis algorithms, we often used a sliding window approach in an attempt to identify a local region in the sequence that may interfere with the process of transcription or translation. In the case of GC content, for instance, the area under the curve and above a threshold was calculated as shown in Figure 1. The largest area and sum of all areas were used to determine whether high or low GC content regions had an influence on successful expression.

Single amino acids and peptides of up to 3 aa that were significantly over or under-represented in each group of proteins ($p < 0.005$) compared to *E. coli* and human proteomes were analyzed (Table 1). Almost every protein is "unusual" to some extent; the question is really to which extent. The set of human proteins expressed here contained over-represented peptides that were rich in isoleucine (I), aspartic acid (D), phenylalanine (F) and glutamic acid (E) compared to the average human protein (Table 1). These over-represented peptides are hydrophilic and flexible. Peptides that were under-represented compared to the average human protein were rich in glycine (G) alanine (A) proline (P) and leucine (L). These peptides have a tendency to be located in coil or turn protein structures and are not charged and not polar. A more uniform bias was seen across all groups when compared to the average *E. coli* protein (Table 1). There was a clear over-representation of peptides rich in serine (S), lysine (K) and glutamic acid (E). These peptides are polar, hydrophilic and tend to occur in coil or turn protein structures. Peptides that were under-represented were rich in alanine (A), leucine (L), and glycine (G) and are non-polar, not charged, not flexible and mostly occurring in α helix protein structures. Interestingly, the amino acid that is over-represented exclusively in group IV compared to the average *E. coli* protein is cysteine (C) (Table 1). This amino acid tends to form disulfide bonds with other cysteines, a modification that cannot occur in the *E. coli* cytosol. Therefore, over-representation of cysteine is usually associated with difficulties expressing cysteine-containing proteins correctly, because the proteins cannot fold correctly. In our set of human proteins, no expression difficulties were observed for proteins rich in cysteine.

Group IV, containing proteins with high expression levels and expected molecular weight, was considered the most optimal for our pipeline. All other groups were compared to it by using an independent sample T-test with assumed equal variance for all sequence analysis methods, followed by one-way ANOVA across all five groups. Decision trees were employed to extract sequence attributes that were the most useful for classification purposes.

Group I contained significantly more hydrophobic proteins than group IV. A negative GRAVY score suggests that a protein is mostly hydrophilic and GRAVY scores were negative for both group I and group IV: -0.21 and -0.42 respectively. However, local hydrophobic regions were shown to be much higher for group I compared to group IV. The average largest hydrophobic region for group I was 15.6, twice the value for group IV. Furthermore, group IV contained a nearly balanced hydrophobic to hydrophilic ratio

A

	E. coli proteome	Human proteome	Group IV
Group I	S, K, SS, C, KKK, SSS, SGG, KK, GS, SK	I, SGG, F, Y, S, KKK, KMR, HI, KMT, HIW	GGG, SGG, WA, S, KKK, FPL, GSS, LSP, IW, F
Group II	S, P, K, PS, E, C, DS, ED, SS, SP	N, YEN, I, EN, AEN, NSD, YRG, DSL, TVD, DS	P, NSD, PT, AY, YEN, VFS, TIT, PSP, NVL, EAQ
Group III	SP, S, ED, EDD, P, SPN, PSP, EED, C, PN	D, SP, EDD, SPN, ED, DD, PN, EED, DPQ, DYE	PSP, SP, SPN, EDD, SDE, LAP, DYE, EED, ED, P
Group IV	C, K, S, E, SS, EE, DHR, SK, KC, PS	DHR, I, D, QYY, N, AN, NYL, EVF, AHY, THH	
Group V	K, E, KD, EE, S, D, NK, KDL, KE, FGP	D, FGP, N, E, KD, AMR, NK, EF, WDE, KDL	FGP, GTV, EVP, VAG, PTS, KVK, IRK, GNK, ALH, AKI

B

	E. coli proteome	Human proteome	Group IV
Group I	A, IG, IA, G, AL, DA, AR, LA, AA, WQ	P, PG, A, PP, GP, KQ, RD, QE, E, CF	E, EV, RD, D, KQ, CF, H, AN, TH, DQL
Group II	A, G, AL, VA, AA, L, M, IA, LI, AI	PG, G, A, GP, AL, C, LLG, RA, RS, PGP	DM, KL, C, GGS, FM, YL, DG, CT, EEL, RF
Group III	A, L, IA, LA, AI, AA, G, GA, AR, M	A, R, L	None
Group IV	A, L, G, AL, AA, AG, VA, M, AR, LL	P, GP, A, G, L, LL, AP, PG, PA, LLL	
Group V	A, LA, G, AA, FA, L, AG, IG, GA, AL	A, P, R, C, LP, RR, G, SS, AP, PS	C, VH, H, LA, GGS, AH, R, SC

Table 1: Top ten over-represented (A) and under-represented peptides in each expression group compared to the average E. coli protein, the average human protein and to the average group IV protein. Peptides were detected using POPPs with probability value of 0.005 and scaling protein length to 100 aa. Peptides are sorted from the most over- or under-represented peptide to the least.

(1.1), while group I was strongly biased towards high hydrophobicity with a ratio of 3.7. Amino acid content of aliphatic, acidic and polar amino acids, which are directly related to hydrophobicity, were all significantly different and correlated to the difference in hydrophobicity (Table 2). Peptides that were over-represented compared to the average human protein and the average protein of group IV had a high hydrophobic content with mean GRAVY values of 0.41 and 0.47 respectively. The average protein complexity score for group I was twice that of group IV (the higher the score the lower the complexity), both for entire protein sequences and for the largest low-complexity regions within them. Furthermore, compared to the average group IV protein, peptides of group I that were over-represented had a very low complexity, such as GGG, SGG and KKK (Table 1). Proteins in group I had a significantly higher isoelectric point and higher charge compared to group IV. The difference in mean between groups was nearly one unit for both isoelectric point and charge. Proteins from group I had a higher β-sheet propensity, 0.24 compared to 0.22 (p=0.06). These proteins were also significantly more rigid, an observation that was also supported by a higher aromaticity score. Proteins from group IV had a lower intrinsic disorder score, 0.07 compared to 0.12 (p=0.008). Intrinsic disorder score, as defined by FoldIndex (Prilusky et al. 2005), was determined by a sliding window analysis of the hydrophobicity plot and by the protein charge. An overall positive score suggested that most of the protein is likely to fold. No significant difference was found when comparing the size of the longest disordered regions between groups or the total number of residues in disorder regions. Codon adaptation index was not significantly different between the two groups, but AAcai localized values suggested that codon usage is slightly less optimal for group I when taking the amino acid content into account. Peptides that were over-represented in group I compared to the average human protein were characterized as basic, bulky, having a high aromatic content and a higher likelihood to occur in β-sheet protein structures. Under-represented peptides were characterized as hydrophilic with very low aliphatic and aromatic content and having a higher likelihood to occur in coil or turn protein structures. Over-represented peptides compared to group IV were slightly hydrophobic, non-polar and had a high aromatic content. Under-represented were hydrophilic, rigid, charged, polar, acidic and more likely to occur in α helix protein structures. Three decision trees were created that classified proteins to group I or IV using 125 proteins in each group. In general more proteins were classified correctly in group IV than in group I (Figure 4A). The best tree classified correctly 73% of the proteins, 85% correctly in group IV and 62% correctly in group I (Figure 4 A2). All three decision trees used the same two top nodes, namely, max region hydrophobic AUC and isoelectric point.

Table 2: Mean values and standard errors of DNA and protein attributes of the five expression groups. P-values on the right column were acquired using a one-way ANOVA test across all five groups. A significant ANOVA p-value ($p < 0.05$) inidcates that the mean across all groups is not equal. The mean values of groups I, II, III and V were further compared to the mean value of group IV using a t-test and significant differences ($p < 0.05$) are indicated in Bold.

Attribute	Description	No bands (group I)	Correct faint bands (group II)	Wrong faint bands (group III)	Correct strong bands (group IV)	Wrong strong bands (group V)	p-Value
DNA length (bp)	length of coding DNA insert	229±7	227±7	**256±18**	221±6	**254±12**	0.03
DNA complexity G1	DNA complexity score calculated using G1	0.16±0.0	0.16±0.0	0.16±0.0	0.16±0.0	0.16±0.0	0.28
DNA complexity SEG	DNA sequence complexity score caculated using nSEG	0.03±0.0	0.03±0.0	0.015±0.01	0.021±0.0	0.03±0.01	0.37
mRNA folding stability	Lowest ΔG of predicted mRNA secondary structure as calculated using mFold	-16.4±0.5	-16.5±0.6	-17±1	-16.9±0.4	-17.2±0.8	0.82
Codon adaptation index (CAI)	Codon adaptation index score calculated using a reference set of 121 highly expressed E. coli proteins	0.39±0.006	0.39±0.007	0.38±0.01	0.39±0.005	**0.41±0.008**	0.12
AA codon adaptation index (AAcai)	CAI calclated while taking amino acid shortage into account	0.26±0.0	0.27±0.0	0.26±0.01	0.27±0.0	0.29±0.01	0.021
Max CAI AUC 0.20	Sum of all area and largest area below CAI or AAcai threshold of 0.2. Values were calculated using the trapezoid method on a plot generated by a sliding window of 4 codons (Figure 2). Other threshold that were tested included: 0.1, 0.15, 0.3, 0.35 and 0.4	0.34±0.02	0.33±0.02	0.32±0.03	0.3±0.01	0.28±0.02	0.26
Total CAI AUC 0.20		**0.77±0.06**	0.7±0.04	0.78±0.09	0.63±0.04	0.63±0.06	0.14
Max AAcai AUC 0.20		0.5±0.03	0.48±0.02	0.45±0.03	0.43±0.02	0.38±0.03	0.03
Total AAcai AUC 0.20		1.3±0.1	1.21±0.1	1.37±0.1	1.1±0.1	1.1±0.1	0.18
GC content	Fraction of GC content in DNA sequence	0.48±0.01	0.49±0.01	0.5±0.01	0.49±0.0	0.48±0.01	0.58
Sum GC content AUC above 65%	See figure 1 for description. Other thresholds that were tested included a GC content of 60% and 70%	1.36±0.2	1.47±0.2	1.66±0.4	1.29±0.1	1.36±0.3	0.84
Max GC AUC above 65%		0.65±0.1	0.73±0.1	0.7±0.2	0.58±0.06	0.64±0.2	0.84
Fraction of GC above 65%		0.09±0.013	0.11±0.013	0.09±0.022	0.1±0.009	0.09±0.018	0.87
Sum AT content AUC above 65%		2.0±0.2	1.85±0.2	1.84±0.4	1.87±0.2	2.48±0.4	0.51
Max AT AUC region 65%		0.95±0.1	0.86±0.1	0.66±0.1	0.87±0.01	1.18±0.2	0.27
Fraction of AT above 65%		0.14±0.01	0.14±0.01	0.11±0.02	0.14±0.012	0.14±0.02	0.91

Attribute	Description	No bands (group I)	Correct faint bands (group II)	Wrong faint bands (group III)	Correct strong bands (group IV)	Wrong strong bands (group V)	p-Value
Protein length (AA)	Length of protein sequence	76±3	75±2.4	**85±6**	73±2	**84±4**	0.03
Molecular weight (Da)	Molecular weight of protein sequence	8579±287	8472±276	**9507±653**	8254±225	**9557±481**	0.03
Aromaticity	aromaticity score calculated according to Lobry and Gautier (Lobry, 1994)	**0.1±0.004**	0.08±0.004	0.08±0.007	0.08±0.003	0.09±0.005	0.01
Protein instability index	protein instability index predicting in-vivo stability of proteins (Guruprasad, 1990). Unstable proteins have values above 40.	45.5±1.96	**48±1.69**	49.5±4.15	43.3±1.28	39.6±2.33	0.025
Average flexibility	Average and maximum flexibility values calculated from the flexibility plot as described by Vinihen et al. (Vihinen, 1994)	**0.99±0.001**	1.003±0.001	**1.006±0.002**	1.002±0.001	1.004±0.001	0.000
Maximum flexibility		1.05±0.001	1.06±0.001	1.06±0.002	1.05±0.001	**1.06±0.002**	0.059
GRAVY	Grand average mean of hydrophobicity as described by Kyte and Doolittle (Kyte, 1982)	**-0.21±0.06**	-0.46±0.04	-0.51±0.07	-0.42±0.03	-0.49±0.06	0.000
Sum hydrophobic AUC	Sum of all hydrophobic/hydrophillic areas and largest hydrophobic/hydrophillic area calculated from a the Kyte and Doollittle hydrophobicity plot (Kyte, 1982) created using a sliding window of 11 aa	**20.9±1.9**	12±1.13	11.9±1.47	11.9±0.93	13.8±1.96	0.00
Max region hydrophobic AUC		**15.6±1.59**	8.5±0.93	7.2±0.93	7.4±0.6	8.9±1.48	0.00
Sum hydrophiliic AUC		36.5±3.4	41.2±2.6	**51.4±6**	37.1±1.9	**48.2±4.5**	0.01
Max region hydrophiliic AUC		26.5±3	28.3±2.1	**36.8±4.9**	26.3±1.7	33.3±3.8	0.14
Normalized hydrophobic AUC	Sum of all hydrophobic/hydrophillic areas and largest hydrophobic/hydrophillic area divided by the sequence length	**0.29±0.03**	0.16±0.02	0.14±0.02	0.16±0.01	0.16±0.02	0.00
Normalized hydrophiliic AUC		0.45±0.03	0.54±0.03	0.61±0.05	0.5±0.02	0.56±0.04	0.02
Hydrophobic/ Hydrophilic	Ratio of hydrophobic AUC to hydrophilic AUC	**3.7±0.99**	1±0.3	0.8±0.4	1.1±0.36	1.7±0.7	0.00
Isoelectric point	Isoelectric point as calculated by pepstats	**7.63±0.22**	7.1±0.22	6.32±0.44	6.74±0.15	**6.13±0.25**	0.00
Protein charge	Charge value as calculated by pepstats	**1.3±0.5**	0.1±0.5	**-3±1.4**	-0.1±0.3	**-2.2±0.6**	0.00
Number of tiny amino acids	Percentage of the amino acids: A+C+G+S+T	28±0.8	26.7±0.6	27.2±1.2	27.7±0.5	26.7±0.8	0.64
Number of small amino acids	Percentage of the amino acids: A+B+C+D+G+N+P+S+T+V	47.9±0.78	49±0.78	50.8±1.5	48.9±0.6	49.3±0.1	0.47

Attribute	Description	No bands (group I)	Correct faint bands (group II)	Wrong faint bands (group III)	Correct strong bands (group IV)	Wrong strong bands (group V)	p-Value
Number of aliphatic amino acids	Percentage of the amino acids: I+L+V	**22.2±0.7**	20.6±0.5	19.8±0.9	20.6±0.4	20.3±0.7	0.08
Number of aromatic amino acids	Percentage of the amino acids: F+H+W+Y	12.2±0.5	11±0.4	10.4±0.7	11.1±0.3	11±0.5	0.19
Number of non-polar amino acids	Percentage of the amino acids: A+C+F+G+I+L+M +P+V+W+ Y	**53.8±0.9**	50.6±0.7	49±1.4	50.6±0.6	49.1±1	0.00
Number of polar amino acids	Percentage of the amino acids: D+E+H+K+N+Q+R+S+T+Z	**46.2±0.9**	49.4±0.7	51±1.4	49.4±0.6	50.9±1	0.00
Number of charged amino acids	Percentage of the amino acids: B+D+E+H+K+R+Z	**24.4±0.7**	27.1±0.8	27.1±1.2	27.4±0.5	28.3±0.9	0.00
Number of basic amino acids	Percentage of the amino acids: H+K+R	13.76±0.5	14.32±0.5	12.48±0.9	14.32±0.4	13.27±0.5	0.20
Number of acidic amino acids	Percentage of the amino acids: B+D+E+Z	**10.7±0.5**	12.8±0.5	14.7±1.2	13.1±0.3	**15±0.6**	0.00
protein helix predicted fraction	Protein secondary structure was predicted using garnier (Garnier, 1978) and the fraction of helix, β-sheet, coil and turn was calculated from the output	0.36±0.02	0.34±0.02	0.33±0.03	0.38±0.02	0.40±0.02	0.36
protein β-sheet predicted fraction		0.24±0.01	0.23±0.01	0.2±0.02	0.22±0.01	0.2±0.01	0.06
protein turn predicted fraction		0.21±0.01	0.22±0.01	0.24±0.02	0.22±0.01	0.2±0.01	0.53
protein coil predicted fraction		0.19±0.01	0.2±0.01	**0.23±0.02**	0.19±0.01	0.21±0.01	0.15
Protein low complexity	Low complexity score and the score of the largest contigous low complexity region calculated using 0j.py (Wise, 2001).	**1.63±0.27**	0.97±0.14	1.26±0.32	0.78±0.1	0.6±0.13	0.00
Protein largest low complexity region		**1.43±0.23**	0.89±0.12	1.11±0.27	0.75±0.09	0.57±0.12	0.002
protein disorder score	Protein disorder, charge and hydrophobic scores calculated using FoldIndex (Prilusky, 2005). From the program output the longest disorder segment and the total number of residues in disorder segments were extracted	**0.12±0.02**	0.05±0.02	0.01±0.03	0.07±0.01	0.04±0.02	0.00
protein disorder charge		**0.053±0.00**	**0.054±0.00**	**0.076±0.01**	0.044±0.00	0.052±0.01	0.00
protein disorder phobic		**0.48±0.01**	0.45±0.01	0.44±0.01	0.45±0.00	0.45±0.01	0.00
protein disorder longest segment		19.7±2.7	**26.1±2.6**	**35.3±6.4**	20±1.8	27.4±4	0.01
number of residues in disorder region		21±3	29±3	**39±8**	23±2	30±4	0.01

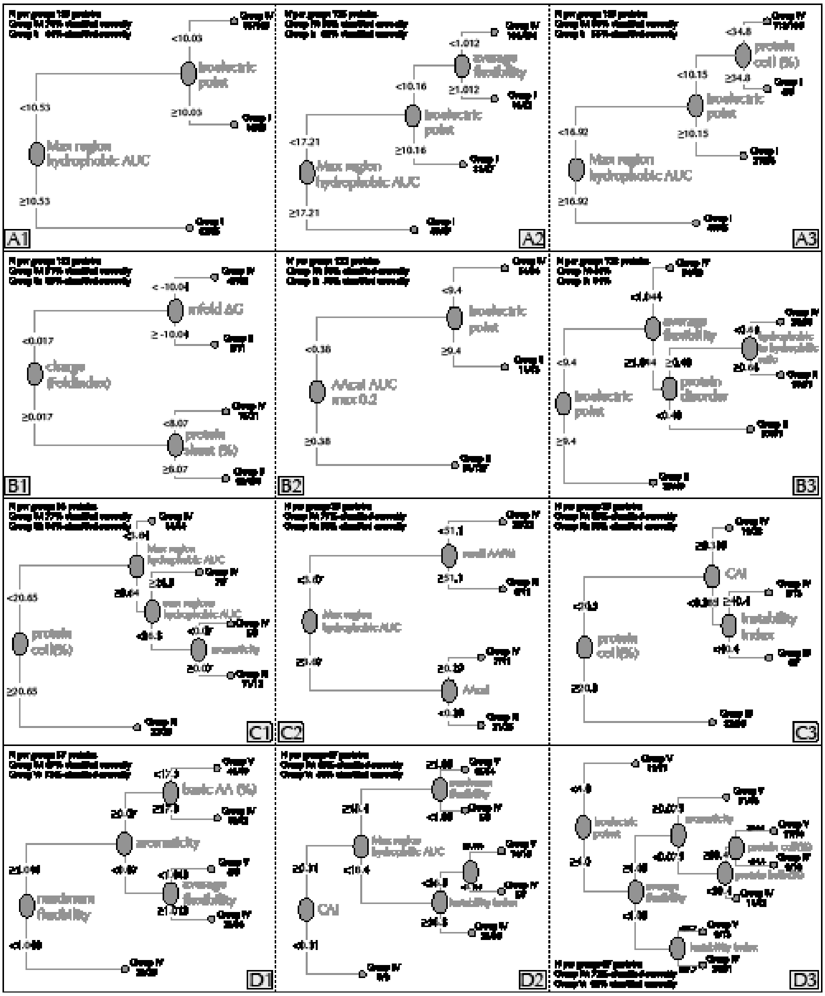


Figure 4: Decision tree classification of each expression group and groups IV (correct strong bands): Group I (no visible bands) and group IV (A); group II (faint correct bands) and group IV (B); group III (wrong faint bands) and group IV (C); group V (wrong strong bands) and group IV (IV). Group IV was the largest expression group in our data set containing 198 proteins. To avoid bias due to unequal group sizes a random selection of proteins was sampled from group IV equal in size to the other group. To increase confidence in the classification process, three trees were constructed, each time sampling randomly from group IV. The length of the branch is proportional to the classification error. Next to each leaf node the predicted group is indicated and the proportion of the number of cases that were classified correctly to the total number of cases predicted for that leaf. For instance, in the lowest leaf of decision tree D1, 29 proteins were classified into group IV, but only 23 of those were actually from group IV and the remaining 6, from group V.

Sequence attributes of group II and group IV were not significantly different for almost all attributes (Table 2). The longest disordered segments were on average longer for proteins in group II. However, the overall number of residues in disorder segments and protein disorder score were not significantly different. Proteins in group II had a higher instability index, 48 compared to 43 for proteins in group IV (an instability value above 40 suggests the protein is unstable). Over-represented peptides, compared to the average group IV protein, were non-polar, small and more likely to occur in coil or turn protein structures (Table 1). Under-represented peptides were slightly hydrophobic and non-polar (Table 1). Three decision trees were created that classified proteins in group II or IV using 122 proteins in each group (Figure 4B). The best decision tree classified correctly 71% of the proteins, 91% correctly in group II and 51% correctly in group IV (Figure 4 B3). Isoelectric point was used in two of the trees and the charge as calculated by FoldIndex in the third.

Group three was the smallest expression group containing 35 proteins. Proteins from this group were significantly more hydrophilic than proteins from group IV. The largest hydrophilic area and the sum of all hydrophilic areas were on average 1.4 times larger for group III (Table 2). No significant difference was observed for the hydrophobic areas. Proteins from group III also had a negative charge of 3, significantly lower than the neutral charge of proteins from group IV. Over-represented peptides, compared to the average protein in group IV, were rich in aspartic (D) and glutamic (E) acids (Table 1). Furthermore, compared to the average human protein, over-represented peptides were rich in aspartic acid (D), glutamic acid (E), serine (S) and proline (P); supporting the observation that group III is characterized mainly by a strong hydrophilic content. There were no significantly under-represented peptides compared to group IV and the only under-represented peptides compared to the average human protein were the single amino acids alanine (A) arginine (R) and leucine (L) (Table 1). Proteins from group III had a slightly higher instability index score ($p=0.08$) and significantly higher disorder. The overall disorder scores were 0.01 and 0.07 ($p=0.054$) for proteins in groups III and IV, respectively. However, the longest disordered segment was 1.75 times longer ($p<0.01$) for proteins in group III and the total number of residues in disordered regions was nearly two fold ($p<0.01$). The protein predicted secondary structure coil propensity was slightly but significantly higher in group

III (4%; $p<0.05$), while protein flexibility was lower than in group IV. Three decision trees were generated that classified proteins to group III or IV using 35 proteins in each group (Figure 4C). The best decision tree classified correctly 86% of the proteins, 94% correctly in group III and 77% correctly in group IV (Figure 4 C1). In two of the three decision trees the coil propensity was used as the initial node. Other attributes related to hydrophobicity and codon usage were also used across the three trees.

Proteins from group V had an average charge of -2.2 and an isoelectric point of 6.13 compared to the neutral charge and isoelectric point of 6.74 for proteins in group IV ($p < 0.01$). Proteins in group V had a higher number of acidic amino acids (15 versus 13, $p < 0.01$) and larger localized hydrophilic regions (48 vs 37, $p < 0.01$). Finally, proteins from group V were slightly but significantly longer than proteins from group IV (11 aa). POPPs detected over-represented peptides that were non-polar with a high aromatic content compared to group IV (Table 1). Peptides that were under-represented were bulky with a high aromatic content. These peptides are characterized as hydrophilic, favorable in helix structures, small and polar. The three decision trees generated with 67 proteins in each group, were each different than the other (Figure 4D). The best decision tree classified correctly 82% of the proteins, 72% correctly in group IV and 92% correctly in group V (Figure 4 D3). Protein flexibility attributes were used across all three trees and protein instability and aromaticity were used in two of the three.

Inclusion body formation

The major attributes that were different between groups were previously associated with inclusion body formation (Idicula-Thomas and Balaji 2005; Baneyx and Mujacic 2004; Ventura 2005). Therefore, 68 human exons were expressed using the same procedure except that proteins were purified separately from the soluble and insoluble phases (Figure 5). Protein expression was visualized on 1D gels and classified into one of five groups as described above. The majority of the protein from the insoluble fractions were visualized as strong bands (groups IV and V), while the majority of the proteins from the soluble fraction were visualized as faint bands (groups II and III) (Figure 6). More proteins that were purified from the insoluble fraction were observed to be of expected size compared to the

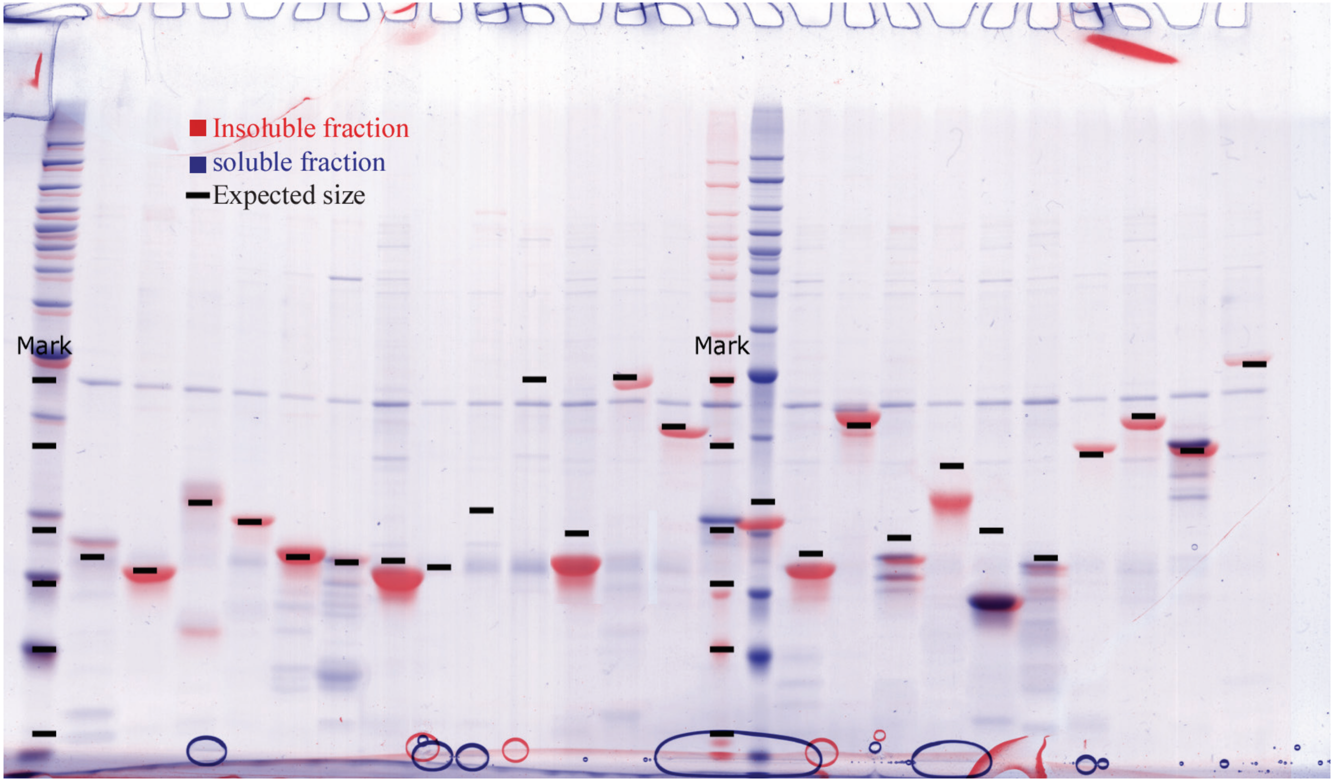


Figure 5: Two 1D-SDS gels from the soluble(blue) and insoluble (red) fractions. The gels were scanned separately and their image was tinted and combined using a graphical image editing program. Each lane contain the exact same sample from both phases, except the right marker lane, which was switched in order to distinguish between the gels. The soluble protein with high molecular weight that is present in all samples is a native E. coli protein that binds to Ni-NTA and should be ignored.

soluble fraction, 43% and 34%, respectively. Furthermore, the number of proteins with no visual band on gel was higher for the soluble fraction. Combining both fractions, no visible protein bands were observed on gel for 7 proteins. However, 37 proteins produced a visible band of expected molecular weight (54%). Of these, 16 were present in both the soluble and insoluble fractions, 13 were present only in the insoluble fraction and 8 were present only in the soluble fraction. Proteins that were expressed correctly in the soluble fraction were compared to proteins that were expressed correctly in the insoluble fraction. The only two parameters that were significantly different were GC content and β-sheet propensity. Proteins that were expressed correctly in the soluble fraction had a higher GC content (61% versus 51%, p = 0.03) and a lower β-sheet propensity (0.11 versus 0.26, p=0.02).

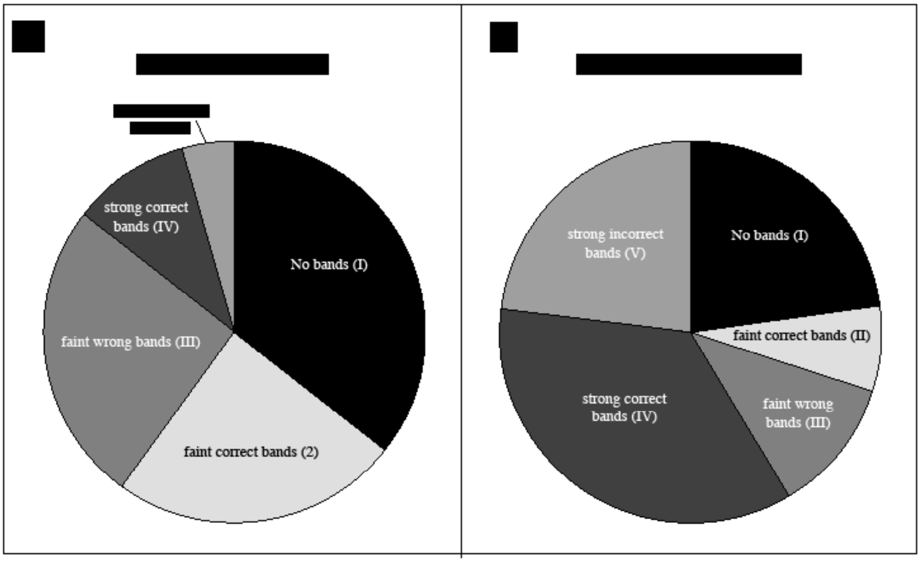


Figure 6: Classification of expressed proteins from the soluble (A) and insoluble fractions (B). Protein purification was performed seperately from the soluble and insoluble fraction of bacterial lysate. Proteins were visualized on 1D-SDS gel and classified into one of five categories: no bands, faint correct/wrong bands, strong correct/wrong bands.

Bacterial growth for these proteins was monitored at an optical density of 550 nm (OD_{550}). The OD_{550} was measured after overnight growth at 30°C, prior to dilution and induction. The average OD_{550} for all 68 proteins was 2.6±0.6. Due to the relatively small number of samples, bacterial growth was compared across three different expression groups, namely, group I (no detectable protein), groups II+IV (expected sizes) and groups III+V (unexpected sizes). A protein was considered correct if the protein band was of the expected size in either of the soluble or insoluble fractions. A protein was classified into group I if no visible band was observed on gel in both the soluble and insoluble fractions. Bacterial growth for group I was significantly lower than the growth observed for groups III+V, which was significantly lower than groups II+IV (Figure 7).

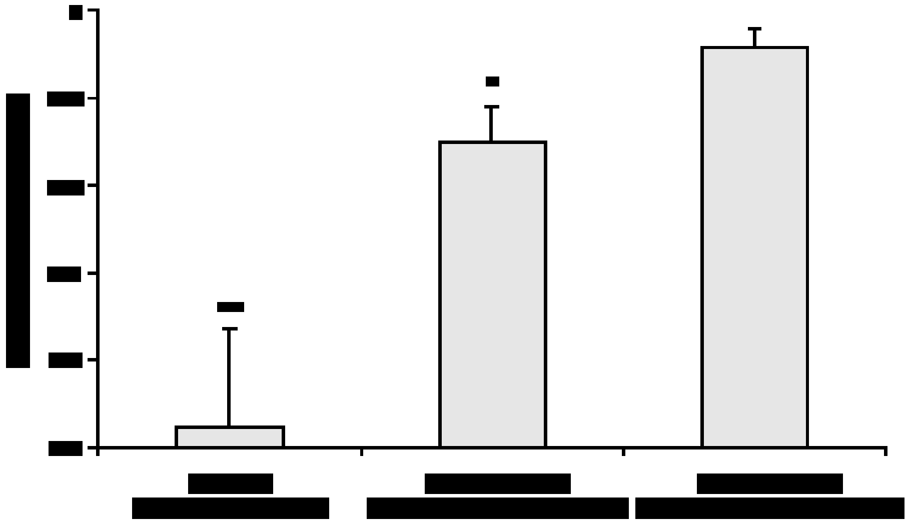


Figure 7: Mean optical density at 550 nm for 68 expressed human exons, after overnight growth at 30°C, prior to dilution and induction. Protein classification into groups was done by considering the combination of the soluble and insoluble fractions. The error bars represent the standard error of the mean value. Group I was significantly smaller (p = 0.002) than groups III+V and II+IV. Group III + V was significantly smaller (p=0.01) than II+IV.

DISCUSSION

The genomics field has been revolutionized by the ability to use high-throughput technology. DNA arrays are now affordable to most labs and genomics information accumulates faster than it can be analyzed. The proteomics field, despite many efforts and advances, still lacks effective high-throughput technology. The work presented here demonstrates the technical difficulties in scaling up heterologous protein expression. We expressed small protein fragments in *E. coli* for generating antibodies against the native proteins. Those fragments were expressed as part of a recombinant protein with a large HisZZ fusion protein on the amino-terminal end and a smaller streptag on the carboxy-terminal end. Although the selected DNA insert was on average a third of the entire recombinant DNA sequence, significant differences were observed in expression level and expression ability in *E. coli*. The success rate reported here of ~60% is similar to previous reports where a pipeline approach was applied (Christendat et al. 2000; Pizza et al. 2000; Braun et al. 2002; Agaton et al. 2003; Luan et al. 2004; Dobrovetsky et al. 2005). All protein bands that were observed on gel were equal or larger to the HisZZ domain size, suggesting this domain is very stable and probably folds independently. We have shown elsewhere that this domain is also beneficial for eliciting antibodies (Zhao et al. 2005).

Part of the difficulty of scaling up protein production is that each protein has distinct physicochemical properties. Naturally, it is possible to optimize the condition for expressing each protein. Unfortunately those conditions are usually found by the "trial and error" approach. Instead of optimizing conditions for each protein, here we attempted to detect, based on sequence analysis, those proteins that are suitable for our specific pipeline protocol. We observed several significant differences between the different expression groups. The group that is most different from all others is group I, which contains the genes whose products failed to elicit visible bands on the gel. The proteins produced by this group were the most hydrophobic, exibited the highest positive charge, highest isoelectric point, highest low-complexity score, lowest flexibility, highest β-sheet propensity and the lowest protein disorder. Hydrophobicity and β-sheet propensity have been previously implicated in the formation of inclusion bodies (Idicula-Thomas and Balaji 2005; Ventura 2005). However, the low flexibility and low protein disorder stand in contrast to inclusion body formation. Inclusion bodies have been shown to be the result of an increased population of partially folded intermediates (Baneyx and Mujacic 2004) and reduced flexibility and disorder are likely to reduce the formation of such intermediates (Ventura 2005). Furthermore, there is a discrepancy between the high positive charge observed in group I, which increases solubility in an aqueous environment, and the large hydrophobic regions which decrease solubility. Therefore, we suggest that the combination of high hydrophobicity and charge together with low flexibility and low protein disorder generates a vulnerable protein that is not likely to form inclusion bodies and is likely to be more efficiently degraded by the bacterial host. This characterization is typical of globular proteins where the non-polar groups are bounded towards the molecule's interior whereas polar groups are bounded outwards, allowing dipole-dipole interactions with the solvent. Proteins that had no visible band exhibited a significantly lower protein complexity and a higher aromaticity score. Low protein complexity is likely to cause amino acid shortage and trigger a stringent response resulting in increased protease activity and protein degradation (Harcum and Bentley 1999; Ramírez and Bentley 1995; Ramírez and Bentley 1999). Ramìrez and Bentley showed experimentally that the addition of phenylalanine to bacteria over-expressing chloramphenicol acetyl-transferase (CAT), reduced the cellular stress and resulted in a proportional increase in production (Ramírez and Bentley 1995). The lower complexity and higher aromaticity score for proteins in group I together with the slower bacterial growth observed, strongly suggests that proteins in this group were produced in smaller quantities. The decision trees distinguished group I from group IV by the single largest hyrdrophobic AUC region and the isoelectric point of the protein, supporting the conclusion that these proteins were probably expressed below detection level due to the combination of protein properties and the heavier burden on the bacterial host.

The two groups of proteins with expected sizes (groups II and IV) were undistinguishable one from the other based on sequence analysis. Despite the relatively high number of proteins in each group, the decision trees generated to classify proteins in one of the two groups were very different one from another with a low correct classification rate, emphasizing further the difficulty of separating these two groups. Final low quantities of purified protein can be explained either by proteolyis, inefficient protein production, inefficient purification or low bacterial growth. Since the ZZ domain was shown to be stable, it is unlikely that the His_6 tag was unavailable for purification. It is also unlikely that bacterial growth was lower for group II since no evidence were found to support such deduction in the other set of 68 proteins where bacterial growth was monitored. POPPs analysis detected over-represented peptides rich in serine (S), a hydrophilic amino acid.

Serine rich peptides were over-expressed in all groups compared to the *E. coli* average protein, however, in group II the over-representation was even higher than in group IV. The solubility experiment showed that soluble proteins were present in low amounts while the insoluble proteins accumulated to larger amounts. Therefore, difference in quantity is most likely due to differences in solubility of the protein or difference in resistance to proteolysis degradation.

Proteins that yielded a visible band on gel but of incorrect size (groups III and V) were either produced as full proteins that were later cleaved or degraded, or the translation simply stopped shortly after the HisZZ domain. Those proteins were more difficult to distinguish from group IV than proteins from group I. The most distinctive property of proteins with incorrect size is their negative charge, compared to the neutral charge of proteins of expected size and higher coil propensity. Those proteins also exhibited larger local hydrophilic regions. Decision trees created to classify proteins were not consistent. However, several attributes that were repeated include percentage of amino acids occurring in coil structures, hydrophobicity and flexibility. Given those properties it is reasonable to assume that the cause for production of proteins with wrong size was either cleavage or degradation of the proteins and not translation interference. However, it is not clear why the amino terminal of those proteins was stable compared to the proteins with no visible bands. The significantly lower bacterial growth for proteins in group I suggests that this group of proteins posed more stress on the bacterial host. Therefore, a resonable explanation is that these proteins were both produced in smaller quantities and were also more prone to complete degradation by the *E. coli* proteases.

Protein expression can fail in any of the different steps, from the stability of the plasmid encoding the protein to the specific protein generated and its interactions with the *E. coli* proteins. Using a pipeline approach, we were not able to determine experimentally whether expression failure occurred at DNA, mRNA or protein levels, only the end result was known. None of the DNA attributes that were derived from sequence analysis methods was significantly different between groups. Therefore, it can be deduced that differences in expression were attributed to the properties of the produced proteins, while the DNA and mRNA constructs were similarly stable for all groups.

A DNA or protein sequence is often characterized using a plot generated by a numerical value assigned to each base or amino acid (Gasteiger et al. 2005), such as the GC content plot shown in Figure 1. Although these plots may be useful when analyzing a few sequences, they are difficult to use when comparing hundred of sequences since there is no easy way to convert them into a number preserving the information displayed. Here we used the area under the curve and above a threshold or the area above the curve and under a threshold depending on the context of the attribute. This method was more useful than using the average value of the plot and emphasized the differences between protein groups. For instance, the GRAVY method as described by Kyte and Doollittle (Kyte and Doolittle 1982) suggested that only the mean GRAVY value of group I was significantly different than from that of group IV while all protein groups were similarly hydrophilic. However, using the mean of the largest hydrophobic and hydrophilic areas in each protein and the mean of sum of all hydrophobic and hydrophilic areas revealed the significantly larger regionalized hydrophilic content in groups III and V compared to group IV and the extent of local hydrophobic regions in group I. Furthermore, attributes that use the area under the curve were shown to be more frequently used in the decision trees compared to attributes that use average values. Therefore, we recommend employing these sequence analysis methods when comparing DNA or protein sequences.

Inclusion body formation was tested on 68 proteins. Clearly, more pure protein product could be obtained from proteins that form inclusion bodies compared to soluble proteins. The majority of the proteins were present in both the soluble and insoluble fractions, however some were present exclusively in one of the two. The exclusive expression in the soluble phase was probably due to the stability and solubility of the HisZZ domain, which is the largest part of the protein. Despite the relatively higher number of proteins with expected size in the insoluble phase, many proteins could clearly be expressed correctly in a soluble form. Therefore, inclusion body formation as such is not essential for correct expression. B-sheet propensity was significantly lower for soluble proteins. Increased β-sheet formations were previously shown to increase the likelihood of inclusion bodies formation (Idicula-Thomas and Balaji 2005; Ventura 2005). The biological significance of the higher percentage of GC content observed in soluble proteins is not clear.

The ability to predict successful expression was limited. Decision trees, which have been previously used for similar purposes (Goh et al. 2003; Bertone et al. 2001) were not consistent except when comparing groups I and IV. While in some cases the same attributes were used in all trees, in many cases, different attributes were used for classification in each tree. This demonstrates the difficulty in determining a small set of attributes that cause expression failure and emphasizes that failure can occur at different levels and due to a different combination of attributes. At this point we are not able to produce an exact algorithm for predicting successful expression with a reasonably good sensitivity and specificity. However, for our specific pipeline, efficiency may be increased simply by avoiding proteins with (i) a strong positive or negative charge (ii) a ratio of hydrophobic AUC to hydrophilic AUC different than 1±0.5 (iii) an isoelectric point below 6.5 or above 7.5 (iv) high aliphatic or aromatic content (v) protein complexity above 1 (vi) high β-sheet or coil content and (vii) low flexibility. Other protocols will be developed to handle those proteins for which specific strategies need to be devised.

Many groups within the academia and industry attempt to produce proteins in a high-throughput manner with varying success rates of 40% to 80%. The success rate is often accepted as is with no further inquiry into the reasons for which protein expression failed for such a large group of proteins. We would like to encourage all groups using high-throughput protein expression to investigate the link between DNA and protein sequences to successful expression. Thereby leading to efficient and affordable protein expression platforms that are essential for proteomics research.

REFERENCES

Agaton C., Galli J., Höidén Guthenberg I., Janzon L., Hansson M., Asplund A., Brundell E., Lindberg S., Ruthberg I., et al. 2003. Affinity proteomics for systematic protein profiling of chromosome 21 gene products in human tissues. *Mol Cell Proteomics* **2:** 405-14.

Baneyx F. and Mujacic M. 2004. Recombinant protein folding and misfolding in Escherichia coli. *Nat Biotechnol* **22:** 1399-408.

Bertone P., Kluger Y., Lan N., Zheng D., Christendat D., Yee A., Edwards A.M., Arrowsmith C.H., Montelione G.T. and Gerstein M. 2001. SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res* **29:** 2884-98.

Braun P., Hu Y., Shen B., Halleck A., Koundinya M., Harlow E. and LaBaer J. 2002. Proteome-scale purification of human proteins from bacteria. *Proc Natl Acad Sci U S A* **99:** 2654-9.

Christendat D., Yee A., Dharamsi A., Kluger Y., Gerstein M., Arrowsmith C.H. and Edwards A.M. 2000. Structural proteomics: prospects for high throughput sample preparation. *Prog Biophys Mol Biol* **73:** 339-45.

Cort J.R., Koonin E.V., Bash P.A. and Kennedy M.A. 1999. A phylogenetic approach to target selection for structural genomics: solution structure of YciH. *Nucleic Acids Res* **27:** 4018-27.

Dobrovetsky E., Lu M.L., Andorn-Broza R., Khutoreskaya G., Bray J.E., Savchenko A., Arrowsmith C.H., Edwards A.M. and Koth C.M. 2005. High-throughput production of prokaryotic membrane proteins. *J Struct Funct Genomics* **6:** 33-50.

Garnier J., Osguthorpe D.J. and Robson B. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* **120:** 97-120.

Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D. and Bairoch A. 2005. Protein identification and analysis tools on the ExPASy server, In *The proteomics protocols handbook* (ed. Walker), pp. 571-607. Humana Press,

Goh C.S., Lan N., Echols N., Douglas S.M., Milburn D., Bertone P., Xiao R., Ma L.C., Zheng D., et al. 2003. SPINE 2: a system for collaborative structural proteomics within a federated database framework. *Nucleic Acids Res* **31:** 2833-8.

Guruprasad K., Reddy B.V. and Pandit M.W. 1990. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng* **4:** 155-61.

Harcum S.W. and Bentley W.E. 1999. Heat-shock and stringent responses have overlapping protease activity in Escherichia coli. Implications for heterologous protein yield. *Appl Biochem Biotechnol* **80:** 23-37.

Harrison R.G. 2000. Expression of soluble heterologous proteins via fusion with NusA protein. *inNovations* **11:** 4-7.

Holm L. and Sander C. 1998. Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics* **14:** 423-9.

Humphery-Smith I. 2004. A human proteome project with a beginning and an end. *Proteomics* **4:** 2519-21.

Idicula-Thomas S. and Balaji P.V. 2005. Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli. *Protein Sci* **14:** 582-92.

Kersey P.J., Duarte J., Williams A., Karavidopoulou Y., Birney E. and Apweiler R. 2004. The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4:** 1985-8.

Kurland C. and Gallant J. 1996. Errors of heterologous protein expression. *Curr Opin Biotechnol* **7:** 489-93.

Kyte J. and Doolittle R.F. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157:** 105-32.

Lobry J.R. and Gautier C. 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. *Nucleic Acids Res* **22:** 3174-80.

Luan C.H., Qiu S., Finley J.B., Carson M., Gray R.J., Huang W., Johnson D., Tsao J., Reboul J., et al. 2004. High-throughput expression of C. elegans proteins. *Genome Res* **14:** 2102-10.

Nilsson B., Moks T., Jansson B., Abrahmsén L., Elmblad A., Holmgren E., Henrichson C., Jones T.A. and Uhlén M. 1987. A synthetic IgG-binding domain based on staphylococcal protein A. *Protein Eng* **1:** 107-13.

Pizza M., Scarlato V., Masignani V., Giuliani M.M., Aricò B., Comanducci M., Jennings G.T., Baldi L., Bartolini E., et al. 2000. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* **287:** 1816-20.

Prilusky J., Felder C.E., Zeev-Ben-Mordehai T., Rydberg E.H., Man O., Beckmann J.S., Silman I. and Sussman J.L. 2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **21:** 3435-8.

Ramírez D.M. and Bentley W.E. 1999. Characterization of stress and protein turnover from protein overexpression in fed-batch E. coli cultures. *J Biotechnol* **71:** 39-58.

Ramírez D.M. and Bentley W.E. 1995. Fed-Batch Feeding and Induction Policies that Improve Foreign Synthesis and Stability by Avoiding Stress Responses. *Biotechnology and bioengineering* **47:** 596-608.

Sharp P.M. and Li W.H. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15:** 1281-95.

Shimada K., Nagano M., Kawai M. and Koga H. 2005. Influences of amino acid features of glutathione S-transferase fusion proteins on their solubility. *Proteomics* **5:** 3859-63.

Skerra A. and Schmidt T.G. 2000. Use of the Strep-Tag and streptavidin for detection and purification of recombinant proteins. *Methods Enzymol* **326:** 271-304.

Ventura S. 2005. Sequence determinants of protein aggregation: tools to increase protein solubility. *Microb Cell Fact* **4:** 11.

Vihinen M., Torkkila E. and Riikonen P. 1994. Accuracy of protein flexibility predictions. *Proteins* **19:** 141-9.

Wan H. and Wootton J.C. 2000. A global compositional complexity measure for biological sequences: AT-rich and GC-rich genomes encode less complex proteins. *Comput Chem* **24:** 71-94.

Wise M.J. 2002. The POPPs: clustering and searching using peptide probability profiles. *Bioinformatics* **18 Suppl 1:** S38-45.

Wise M.J. 2001. 0j.py: a software tool for low complexity proteins and protein domains. *Bioinformatics* **17 Suppl 1:** S288-95.

Wootton J.C. and Federhen S. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* **266:** 554-71.

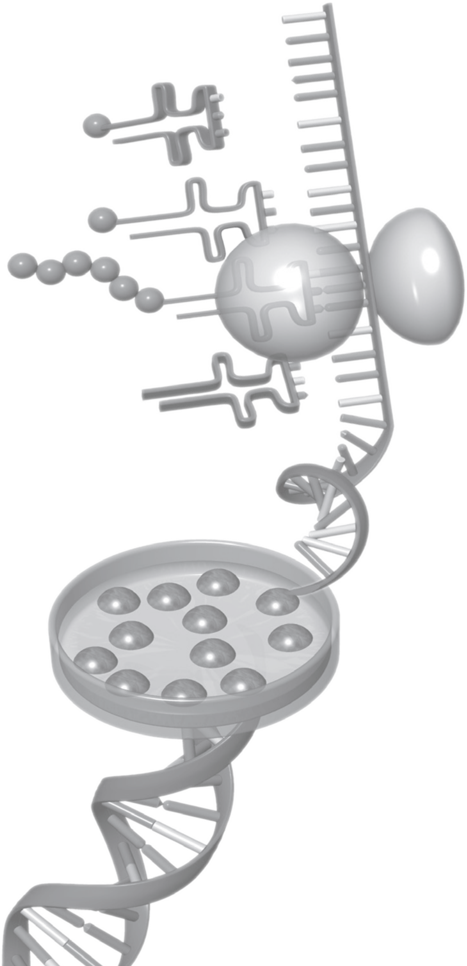
Zarembinski T.I., Hung L.W., Mueller-Dieckmann H.J., Kim K.K., Yokota H., Kim R. and Kim S.H. 1998. Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proc Natl Acad Sci U S A* **95:** 15189-93.

Zhao Y., Benita Y., Lok M., Kuipers B., van der Ley P., Jiskoot W., Hennink W.E., Crommelin D.J. and Oosting R.S. 2005. Multi-antigen immunization using IgG binding domain ZZ as carrier. *Vaccine* **23:** 5082-90.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31:** 3406-15.

7

Summary and Future Perspectives



The great challenge of the post-genomic era is to define and annotate the proteome which is responsible for the bulk of molecular work in the cell. The human proteome is far more daunting in size and complexity than the genome and faces many technical and financial challenges. Proteome research can be divided to two separate phases, the identification of all proteins encoded by the genome and the development of an experimental platform to query and browse the protein content of the cell (e.g. protein- and antibody arrays). In this thesis challenges in both of these aspects have been confronted.

The identification of the human proteome is pursued on the bioinformatics, genomics and proteomics fronts. Advances in gene prediction programs and accumulation of data on mRNA expression in various tissues have contributed significantly to the identification of proteins encoded by the genome. Proteomics research has also seen significant advances in the ability to separate, identify and characterize proteins from a crude mixture of cell extracts. However, each field carries its own limitations and some protein classes fall in the gray area of the mentioned research approaches and are systematically ignored. One such class is the class of small protein-coding genes and the challenges in their identification and a strategy to discover new small genes is discussed in **part I** of this thesis. Small genes are defined as genes coding for a protein smaller than 100 amino acids. Many of the small annotated human genes are involved in signal transduction, inflammation and immune response; systems that are of great interest to the scientific and pharmaceutical communities.

Since the early 1990's genome annotators established the 100 codons threshold as a detectable limit for protein-coding genes. Although it was an accepted compromise at that time, this threshold is still applied today to increase confidence that a predicted gene is real. In **chapter 2** a detailed overview of currently annotated small human genes was presented. This chapter established the existence of small genes in the human genome, showing that small proteins can be encoded directly by small genes as opposed to small proteins that are the cleavage products of high molecular weight proteins. A thorough literature review combined with our own bioinformatics analyses was presented that outlined the challenges of identifying those genes both computationally and biologically. These can be summarized as follows: (i) the small number of codons and amino acids available for statistical analysis in small genes limits the efficiency of gene prediction programs; (ii) the vast number of pseudogenes in the human genome, most of which have 1 to 3 exons, introduce difficulties in identifying small genes; (iii) random sequencing of mRNA from tissue is biased towards highly expressed long genes; (iv) small proteins are inherently difficult to study due to their potentially low intracellular concentration, poor solubility, high diffusibility during electrophoresis and/or column chromatography and the low numbers of amino groups and other dye/isotope accepting elements. The genomic structure of small annotated genes and their annotated function were further characterized in chapter 2. The majority of small genes had 2 or 3 exons, up to 7 and most of the coding sequence was contained within one exon. On average exons of small genes were half the size of exons of large genes, emphasizing the difficulty in locating many of these exons. Gene prediction programs were shown to have a reduced sensitivity but a high specificity in identifying small genes, suggesting prediction programs are optimized to detect fewer genes with a higher level of confidence at the expense of missing real genes. Protein annotation revealed that several important families were in fact a family of small genes. These include proteins involved in signal transduction, inflammation and immune response. Interestingly a large proportion of these proteins were discovered from the protein end, emphasizing once more the bias against identifying small genes from the genome end.

Developers and users of gene prediction programs often ignore small predicted genes in order to avoid a large number of false positives. In **chapter 3**, instead of filtering predicted genes based on their size, we have generated a more sensitive algorithms biased toward identification of small protein-coding genes. The gene structure, genomic location and protein attributes were used to increase confidence in the predicted proteins. The algorithm developed was based on a combination of gene prediction programs, homology to expressed sequence tags (ESTs), homology to known proteins and protein motifs and conserved evolutionary elements. The algorithm generated many more predicted exons than an equivalent algorithm trained only on large genes. One of the difficulties in predicting small genes is the vast number of pseudogenes (genes that lost the ability to encode a functional protein). The predicted small genes were compared to annotated pseudogenes and were shown to avoid most of them. The likelihood of a predicted gene being a pseudogene was further decreased by avoiding genes with in-frame stop codons and genes with self homology (occurring elsewhere in the human genome). Predicted genes were classified into several sets with varying degrees of confidence. In total 1,665 multi-exon and 7,709 single exon small genes were predicted in the human genome. Many more non-annotated predicted small genes were located within introns of annotated genes than in intergenic region, probably due to non-annotated alternative splicing exons. Although many more predicted small genes were single exon genes, we believe these are in-fact exons of small genes that were more easily detected than others. Therefore, a biological approach to validate those genes was suggested and is essential to complete the annotation of small genes.

Protein production is an essential step in many aspects of proteomics and is the most basic requirement in the development of an experimental high-throughput platform to query and browse the protein content of a cell. In **part II** of this thesis barriers of high-throughput protein production were investigated . The major problem in protein production is that proteins differ in their physicochemical properties and no single protocol can be expected to work for all proteins. While it is possible to optimize a protocol to express a specific protein it is not feasible on a large scale when expressing hundreds of proteins in parallel. Therefore, we attempted to correlate protein properties, derived from sequence alone, to their propensity to be successfully expressed under a specific protocol. Thereby, avoiding *a priori* expression of proteins with a low success probability and eventually increasing the pipeline efficiency.

In **chapter 4** a detailed overview was presented discussing the potential problems that may influence high-throughput protein production from the initial step of PCR to the final stages of protein purification. Problems that were discussed at PCR level included primer design considerations and analysis of the PCR template to be amplified. Problems raised with respect to expression in *E. coli* included biased codon usage, biased amino acid content, inclusion body formation and proteolysis. In the next two chapters those problems discussed in chapter 4 were tested on a high-throughput protein production platform.

In **chapter 5** a pipeline approach was applied to amplify a set of 1,438 human exons by PCR. The success rate was 83% and DNA sequence analysis was performed to identify sequence properties that correlate with PCR failure. The main findings of this work suggested that primer design consideration had little effect on success or failure of PCR. Rather the GC content of the template was the most significant property affecting the success rate. An algorithm was developed that quantified local GC content to estimate the probability of PCR success. Those sequences with high local GC content should be avoided in our high-throughput platform and another more suitable protocol should be developed.

In **chapter 6** a set of several hundred proteins was expressed in *E. coli*. The overall success rate was ~60% and was consistent with reports from other groups (Christendat et al. 2000; Pizza et al. 2000; Braun et al. 2002; Agaton et al. 2003; Luan et al. 2004; Dobrovetsky et al. 2005). Proteins sequence analysis algorithms were applied to the set of expressed proteins and were correlated to successful expression. While the results were not clear enough to allow the generation of a reliable algorithm that predicted success probability, several observations were made that allow increasing the efficiency of the production pipeline. These included avoiding highly negatively or positively charged proteins; avoiding hydrophobic to hydrophilic ratio different than 1±0.5; avoiding proteins with high aliphatic and aromatic content and avoiding proteins with a high beta-sheet or coil content. Successful protein expression was previously associated to the ability of expressing proteins in inclusion bodies (Fahnert et al. 2004). Therefore, the link between successful expression and inclusion body formation was investigated by purifying a set of 68 proteins from the soluble and insoluble fractions, separately. The results suggested that most proteins were expressed in inclusion bodies but inclusion body formation was not a requirement for successful expression. Several proteins were shown to be successfully expressed and be exclusively present in the soluble fraction.

FUTURE PERSPECTIVES

In this thesis two aspects that form gaps in current proteomics research have been addressed. The identification of small protein-coding genes and the analysis of barriers in high-throughput protein production in *E. coli*. The work presented on small protein-coding genes is an initial attempt to look into this class of genes. Small genes have been systematically ignored both by the genomics and proteomics communities. This was acceptable at the early stages of genome annotation but considering the accumulation of data and the current status of genome annotation, these genes must now be specifically targeted and identified. In part I of the thesis an extensive work was done to identify small genes despite all the challenges associated with them. It is essential to validate these predicted genes biologically and to sequence their entire mRNA. Because of their small size, we expect that many of the small proteins will have regulatory functions in the cell. Novel validated small genes have to be further studied and characterized and their function in health and diseases has to be established. Furthermore, this study was conducted on the human genome alone but small genes will probably also be present in other species. The identification of small genes in other (model) organism and the subsequent over-expression or knocking out the expression of such genes may potentially lead to the development of interesting animal models. Small proteins may become the drug targets of the future. The extent of small genes in human and other species is still unclear, however, results presented here suggest that these may exist in far greater numbers than previously estimated. Small genes have been primarily studied in bacteria and yeast, for which most genes are contiguous, and were thought to be present in very small numbers in humans. As more small genes will be identified it will also be possible to track the evolution of homologous small genes through several species and determine if there is a selective pressure against small genes.

Many different protein motifs were identified in the predicted small proteins. These motifs should be further studied and their specific annotation may provide more insight into the likelihood that a small predicted gene is real. For instance, a motif of a conserved low complexity protein that occurs in plants may decrease the likelihood that the predicted small protein is real (although one can never be sure) but another motif occurring in a subunit of an RNA processing protein may increase the likelihood. The significance of the different motifs is still not clear at this point and should be carefully sorted and analyzed.

The high-throughput production of proteins is one of the greatest challenges facing proteomics. Overcoming this barrier will launch proteomics into the next phase enabling the efficient development of protein arrays and proteins for generation and selection of antibodies. One protocol for high-throughput protein expression cannot be expected to be adequate for most of the proteins. Instead, several protocols have to be established and each protein should first be produced in the protocol for which the success probability is best. We have attempted this approach by testing the expression of 617 proteins using one protocol. However, since a pipeline approach was used, it was not possible to determine experimentally the cause for expression failure, only the end result was known. A protein can fail to be expressed due to many different reasons, a few examples include: low copy number of the plasmid, instability of the mRNA, incompatible codon usage and toxicity of the produced protein. The cause for failure could be related to the DNA, RNA or protein sequence. The process of protein expression can be broken down to several sub processes: (i) stability and half-life of the plasmid; (ii) stability and half-life of the mRNA; (iii) translation efficiency; and (iv) protein behavior in the bacterial host. A model for each step should be constructed and the sum of all models should give a better result than one model attempting to combine all these steps into one. In order for efficient algorithms to be developed that predict successful protein expression, it is important to determine the cause of failure and attempt to correlate it with the appropriate sequence. In parallel, I suggest to create a database for protein expression and their protocols. The Gerstein group in Yale made the SPINE database, a system for structural proteomics (Bertone et al. 2001; Goh et al. 2003). This system is used to keep track of cloning, expression and purification of proteins for structural proteomics. Similarly, a database for high-throughput protein expression and their associated experimental protocols should be made. Over a certain period of time enough data will accumulate of proteins that were successfully expressed with one protocol or failed with another. Users of the database will be able to search, based on sequence homology for similar proteins in the database and retrieve a set of homologous sequences that were successfully expressed or failed to be expressed with a specific protocol. In such a manner a set of hundreds of proteins could be used to query the database and retrieve the smallest number of protocols to express the protein set. Each protein will be expressed with a protocol that was previously applied successfully. Such a database can only be successful if adopted by several groups interested in large scale protein expression. The number of proteins to be expressed on a proteomic scale may be immense but it is a finite number of proteins. With time such a database could prove to be a very valuable tool for protein expression and could be used for data mining and for generation of better prediction algorithms.

REFERENCES

Agaton C., Galli J., Höidén Guthenberg I., Janzon L., Hansson M., Asplund A., Brundell E., Lindberg S., Ruthberg I., et al. 2003. Affinity proteomics for systematic protein profiling of chromosome 21 gene products in human tissues. *Mol Cell Proteomics* 2: 405-14.

Bertone P., Kluger Y., Lan N., Zheng D., Christendat D., Yee A., Edwards A.M., Arrowsmith C.H., Montelione G.T. and Gerstein M. 2001. SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res* **29:** 2884-98.

Braun P., Hu Y., Shen B., Halleck A., Koundinya M., Harlow E. and LaBaer J. 2002. Proteome-scale purification of human proteins from bacteria. *Proc Natl Acad Sci U S A* **99:** 2654-9.

Christendat D., Yee A., Dharamsi A., Kluger Y., Gerstein M., Arrowsmith C.H. and Edwards A.M. 2000. Structural proteomics: prospects for high throughput sample preparation. *Prog Biophys Mol Biol* **73:** 339-45.

Dobrovetsky E., Lu M.L., Andorn-Broza R., Khutoreskaya G., Bray J.E., Savchenko A., Arrowsmith C.H., Edwards A.M. and Koth C.M. 2005. High-throughput production of prokaryotic membrane proteins. *J Struct Funct Genomics* **6:** 33-50.

Fahnert B., Lilie H. and Neubauer P. 2004. Inclusion bodies: formation and utilisation. *Adv Biochem Eng Biotechnol* **89:** 93-142.

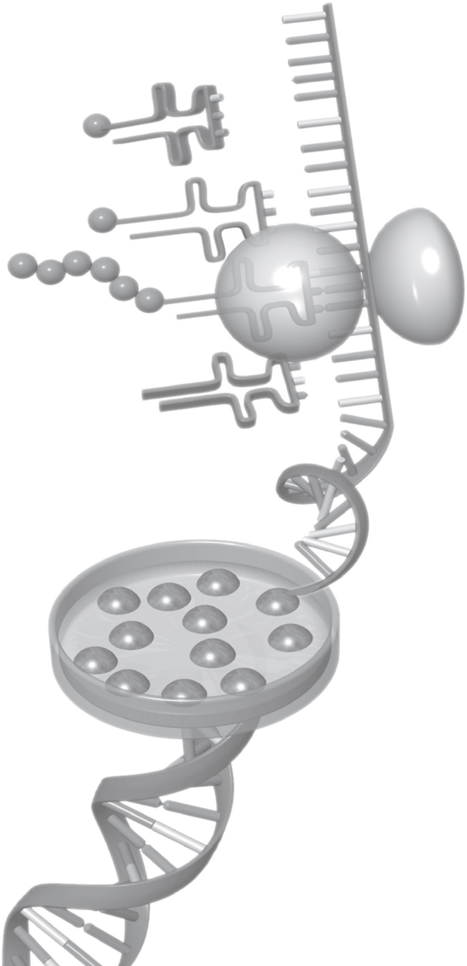
Goh C.S., Lan N., Echols N., Douglas S.M., Milburn D., Bertone P., Xiao R., Ma L.C., Zheng D., et al. 2003. SPINE 2: a system for collaborative structural proteomics within a federated database framework. *Nucleic Acids Res* **31:** 2833-8.

Luan C.H., Qiu S., Finley J.B., Carson M., Gray R.J., Huang W., Johnson D., Tsao J., Reboul J., et al. 2004. High-throughput expression of C. elegans proteins. *Genome Res* **14:** 2102-10.

Pizza M., Scarlato V., Masignani V., Giuliani M.M., Aricò B., Comanducci M., Jennings G.T., Baldi L., Bartolini E., et al. 2000. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* **287:** 1816-20.

8

Samenvatting



Eén van de grootste uitdagingen in de "life sciences" is de karakterisatie van het proteoom, de totale verzameling van eiwitten die in een organisme voorkomt. Nu het humane genoom bekend is, is de volgende stap het vaststellen voor welke eiwitten dit genoom codeert en wat de functie is van deze eiwitten. Het humane proteoom is zeer groot en complex en daardoor ook zeer kostbaar om te bestuderen. Het is daarom ook van belang dat er nieuwe bioinformatica-benaderingen en experimentele "high-throughput" technieken (zoals eiwit- en antilichaamarrays) worden ontwikkeld om het humane proteoom te karakteriseren. In het onderzoek dat in dit proefschrift is beschreven is aan beide aspecten gewerkt.

Bij de identificatie van het humane proteoom wordt gebruikgemaakt van bioinformatica, genomics en proteomics. De nieuwe generatie genvoorspellingsprogram ma's gecombineerd met de grote hoeveelheid aanwezige informatie over mRNA expressie (met name expressed sequence tags, ESTs) in zeer veel verschillende weefsels en cellen hebben significant bijgedragen aan de identificatie van de genen die coderen voor eiwitten. Daarnaast zijn de huidige technieken in het proteomics-onderzoek sterk verbeterd en maken identificatie van eiwitten in complexe biologische monsters mogelijk.

Echter bepaalde klassen van eiwitten kunnen tot nu toe niet goed in kaart worden gebracht. Het gaat dan met name om eiwitten kleiner dan 100 aminozuren. In **deel I** van dit proefschrift wordt ingegaan op de specifieke problemen verbonden aan de identificatie van deze klasse van eiwitten en wordt een bioinformatica-strategie gepresenteerd om nieuwe kleine eiwitten te ontdekken. Er zijn op dit moment enkele honderden kleine eiwitten in de mens bekend en deze hebben een functie in bijvoorbeeld signaaltransductie en in de ontsteking/immuun response.

Tot nu toe wordt voor de annotatie van genen in een genoom een grens aangehouden van meer dan 100 codons voor de detectie van genen die coderen voor eiwitten. Deze grens wordt aangehouden om met voldoende zekerheid te kunnen zeggen of een voorspeld gen ook een echt gen is. In **hoofdstuk 2** wordt een overzicht gegeven van alle tot nu toe geannoteerde kleine genen in het humane genoom. Veel kleine eiwitten, die in de mens voorkomen, zijn gemaakt als een onderdeel van een veel groter eiwit en door de activiteit van proteolytische enzymen wordt een kleiner, biologisch actief eiwit hieruit gevormd. Daarnaast zijn er dus kleine eiwitten die direct worden gecodeerd door een klein gen. In hetzelfde hoofdstuk wordt een literatuuroverzicht en een door ons zelf uitgevoerde bioinformatica-analyse gepresenteerd waaruit duidelijk na voren komt welke problemen verbonden zijn aan het identificeren van dergelijke kleine eiwitten/genen. Deze problemen liggen zowel op het vlak van de bioinformatica als in het laboratorium. De belangrijkste problemen zijn: (i) het kleine aantal codons/aminozuren dat beschikbaar is voor de statistische onderbouwing veroorzaakt dat de voorspellingen door genvoors pellingsprogramma's erg onzeker zijn; (ii) het grote aantal pseudogenen in het humane genoom, met gemiddeld dezelfde hoeveel exonen als een klein gen, bemoeilijkt de identificatie van kleine genen; (iii) in experimenten waarin random mRNAs/cDNAs worden gesequenced, worden vooral mRNAs bepaald van grotere eiwitten die relatief hoog tot expressie komen; (iv) kleine eiwitten zijn moeilijk te bestuderen doordat ze potentieel in kleine hoeveelheden voorkomen, mogelijk niet goed oplossen, een hoge diffussiesnelheid hebben in chromatografische scheidingstechnieken en moeilijk te labelen (b.v. fluorescent) zijn door het kleine aantal aanwezige reactieve groepen in het betreffende molecuul. De genstructuur van kleine bekende genen en hun functie worden ook bediscussieerd in hoofdstuk 2. De meeste bekende kleine genen hebben 2 tot 3 exonen, met een maximum van 7. De coderende informatie bevindt zich voor het grootste deel in maar één exon. Gemiddeld zijn de exons van kleine genen half zo groot als die van grotere genen. Dit onderstreept nog eens waarom het zo moeilijk is om kleine genen te herkennen in het genoom.

De beschikbare genvoorspellingsprogramma's zijn minder gevoelig voor het vinden van reeds geannoteerde kleine genen dan voor het vinden van geannoteerde grote genen. De specificiteit van deze programma's voor het vinden van een klein gen is echter wel hoog. Deze programma's zijn dus geoptimaliseerd voor het vinden van genen met een grote zekerheid, maar dit heeft tot gevolg dat deze programma's niet alle aanwezige genen kunnen identificeren. Uit de analyse van de bekende kleine eiwitten bleek verder dat veel van deze eiwitten zijn ontdekt vanuit het eiwit en dus niet vanuit het DNA.

Ontwikkelaars en gebruikers van genvoorspellingsprogramma's verwijderen kleine genen uit de resulaten omdat ze bang zijn voor een groot aantal valspositieven. In **hoofdstuk 3** wordt een door ons ontwikkeld algoritme gebruikt voor het vinden van genen die coderen voor kleine eiwitten. Genstructuur, de locatie in het genoom en eigenschappen van het voorspelde eiwit (b.v. het voorkomen van een eiwitmotief) worden hierin gebruikt om met een grotere zekerheid kleine genen te kunnen voorspellen. Het ontwikkelde algoritme maakt gebruik van een combinatie van bestaande genvoorspellingsprogramma 's, homologie met "expressed sequence tags" (EST), homologie met bekende eiwitten en eiwitmotieven en de aanwezigheid van evolutionair geconserveerde elementen in het gen. Het algoritme voorspelt veel meer exonen dan een vergelijkbaar algoritme voor het vinden van genen van grotere eiwitten. Eén van de grootste problemen in het voorspellen van kleine genen is het grote aantal pseudogenen in het humane genoom. Pseudogenen coderen niet voor een functioneel eiwit, b.v. doordat ergens in de sequentie een stopcodon is ontstaan. De voorspelde kleine genen werden daarom in eerste instantie vergeleken met geannoteerde pseudogenen. Alle hits werden verwijderd uit de voorspelde verzameling van kleine genen. De kans dat een voorspeld klein gen een pseudogen is, werd verder verkleind door genen uit deze verzameling te verwijderen met "in frame" stopcodons of met homologie ergens anders in het genoom. In totaal werden 1665 multi-exon en 7709 één-exon kleine genen voorspeld in het humane genoom. De meeste van deze niet-geannoteerde kleine genen werden gevonden in introns van geannoteerde grote genen en niet zo zeer in het DNA tussen genen. Het zou kunnen dat veel van de kleine genen in introns feitelijk niet-geannoteerde exonen zijn van het omliggende gen. Van de meeste voorspelde kleine genen is maar één exon voorspeld. Waarschijnlijk hebben we voor deze genen in de meeste gevallen één of meerdere exonen gemist in onze analyse. Het is daarom ook van groot belang dat onze bioinformatica resultaten worden gevalideerd op biologsich materiaal. Met andere woorden: komen de voorspelde kleine genen ook echt voor?

High-throughput recombinant eiwitproductie is een belangrijke stap in de ontwikkeling van high-throughput test systemen (eiwit en antilichaam arrays) voor het bestuderen van het proteoom. In **deel II** van dit proefschrift worden experimenten beschreven waarin problemen en mogelijk oplossingen verbonden aan high-throughput eiwitexpressie in *E. coli* werden onderzocht. Het belangrijkste probleem in recombinant eiwitproductie is dat eiwitten verschillen in fysisch-chemische eigenschappen. Elk eiwit is uniek, waardoor één bepaald protocol waarschijnlijk niet zal werken voor alle eiwitten. We hebben geprobeerd om eigenschappen van individuele eiwitten (deze eigenschappen werden “berekend” op basis van de DNA- of eiwit sequentie) te correleren aan hoe succesvol zo’n eiwit gemaakt kan worden in *E. coli* volgens een high-throughput protocol. In **hoofdstuk 4** wordt een overzicht gegeven van alle problemen verbonden aan high-throughput eiwit clonering en expressie. Problemen met de initiële PCR-reacties tot en met de eiwitzuiveringmethode worden besproken. In hoofdstuk 5 en 6 worden twee van de problemen die in hoofdstuk 4 werden besproken verder onderzocht. In **hoofdstuk 5** werden de resulaten van 1483 PCR-reacties op humaan DNA geanalyseerd. In 83% van de gevallen was de PCR succesvol verlopen. Op basis van de DNA-sequentie is onderzocht waarom de PCR in sommige gevallen niet lukte. Uit de analyse bleek dat primer design niet echt belangrijk was voor een succesvolle PCR. De hoeveelheid guanine (G) en cytosine (C) nucleotiden in het te amplificeren DNA bleek veel belangrijker. Er is een algoritme ontwikkeld dat op basis van lokale hoeveelheden G en C’s voorspelt of een PCR-reactie zal lukken. In high-throughput PCR-reacties zouden dergelijke DNA-sequenties met lokaal veel G en C’s moeten worden vermeden.

In **hoofdstuk 6** worden de resultaten geanalyseerd van een experiment waarin enige honderden eiwitten in *E. coli* tot expressie zijn gebracht. Ongeveer 60% van de in *E. coli* tot expressie gebrachte eiwitten konden uiteindelijk ook worden geïsoleerd. Met behulp van verschillende sequentie-analyse algoritmen is geprobeerd om succesvolle eiwitexpressie te correleren aan eigenschappen van het recombinante eiwit. De resultaten waren helaas niet zodanig dat een betrouwbaar algoritme kon worden ontwikkeld. Echter op basis van de resulaten kunnen wel aanbevelingen voor succesvolle high-throughput heterologe eiwitsynthese in *E. coli* worden gedaan: sterk geladen (positief of negatief) eiwitten moeten worden vermeden; de ratio tussen hydrofobe en hydrofiele delen moet liggen tussen 1 ± 0.5; eiwitten met veel alifatische of aromatische aminozuren of met veel “beta sheet” of “coil” structuur zouden moeten worden vermeden. In hoofdstuk 6 wordt verder nog onderzoek beschreven waarin de vraag centraal stond of voor succesvolle expressie van een heteroloog eiwit in *E. coli*, de vorming van “inclusion bodies” belangrijk is. Hiervoor werden 68 eiwitten geïsoleerd zowel uit de oplosbare als uit de onoplosbare fractie (inclusion bodies). Hoewel de meeste eiwitten werden geproduceerd in de onoplosbare fractie was de vorming van “inclusion bodies” geen voorwaarde voor succesvolle heterologe expressie. Een aantal van de eiwitten werd namelijk alleen maar teruggevonden in de oplosbare fractie.

In dit proefschrift zijn twee heel verschillende problemen in het huidige proteomicsonderzoek onderzocht en is geprobeerd om met behulp van bioinformatica een oplossing voor deze problemen te vinden. Het eerste probleem is dat er mogelijk meer eiwitten bestaan dan dat we op dit moment hebben aangetoond, waarbij met name zou gaan om kleine eiwitten. Het tweede probleem dat onderzocht is, is de high-throughput cloning en expressie van heterologe eiwitten in *E. coli*. Voor beide problemen zijn *in silico* oplossingen gevonden. Uit vervolg onderzoek zal moeten blijken hoe waardevol deze oplossingen in de praktijk zullen zijn.

List of Abbreviations

Publications

Curriculum Vitae

Acknowledgements

LIST OF ABBREVIATIONS

aa	amino acids
B. Subtilis	*Bacillus subtilis*
bp	base-pair
CAT	chloramphenicol acetyltrasferase
CCP	clustered conserved and predicted sequences
cDNA	complementary DNA
DNA	deoxyribose nucleic acid
E. coli	*Escherichia coli*
EST	expressed sequence tag
Gb1	streptococcal protein G
GST	glutathione S-transferase
His6	hitidine tag containing a repeat of 6 histidines
HT	high-throughput
IPTG	isopropyl-beta-D-thiogalactopyranoside
kb	kilo base
kDa	kilo dalton
MBP	maltose binding protein
mRNA	messenger RNA
NusA	N utilization substance protein A
ORF	open reading frame
PCR	polymerase chain reaction
RNA	ribonucleic acid
S. cerevisiae	*Saccharomyces cerevisiae*

PUBLICATIONS

Benita, Y., Oosting, R.S., Lok, M.C., Wise, M.J. and Humphery-Smith, I. (2003). Regionalized GC content of template DNA as a predictor of PCR success. *Nucleic Acids Res*. **31**(16): e99, 1-7.

Zhao, Y., Benita, Y., Lok, M.C., Kuipers, B., Ley, P., Jiskoot, W., Hennink, W.E., Crommelin, D.J.A. and Oosting, R.S. (2005). Multi-antigen immunization using IgG binding domain ZZ as carrier. *Vaccine*. **23**(43):5082-90.

Benita, Y., Oosting, R. S., Xuan, Z., Wise, M. J., Zhang, Q. M. and Humphery-Smith, I. (2006). Challenges in the prediction of small genes in the human genome. *Submitted*.

Benita, Y., Oosting, R. S., Xuan, Z., Wise, M. J., Zhang, Q. M. and Humphery-Smith, I. (2006). Prediction of small protein-coding genes in the human genome. *In preparation*.

Benita, Y., Wise, M. J., Lok, M. C., Humphery-Smith, I., and Oosting, R. S. (2006). Analysis of high-throughput protein expression in *Escherichia Coli*. *Submitted*.

CURRICULUM VITAE

Yair Benita was born on May 18th 1973 in Jerusalem, Israel. In 1991 he completed his high-school diploma with specialization in chemistry. That year he joined the Israel Defense Forces where he initiated and designed computer simulators for training armor officers in navigation, tactics and combat. In October 1995 he started his pharmacy education at the School of Pharmacy in the Hebrew University of Jerusalem, Israel. During his studies he worked as an undergraduate research assistant in Prof. S. Benita's lab where he developed and characterized emulsions and semi-solid drug delivery systems. In addition, he completed a research project investigating the effects of UV radiation on antioxidants secretion from rat's skin in the department of Pharmaceutics in Dr. R. Kohen's lab. In 1999 he performed a six months pharmacy stage in Oplatka pharmacy in Jerusalem and in April 2000 he graduated and obtained his pharmacist license from the Israeli minister of health. In October 2000 he started his PhD project in bioinformatics under the supervision of Prof. Ian Humphery-Smith and Dr. Ronald Oosting in the department of Pharmaceutical Proteomics in Utrecht University, The Netherlands. During his PhD period he attended several international conferences, including ISMB (Intelligent Systems for Molecular Biology; Denmark 2001 and Scotland 2004) and "Identification of Functional Elements in Mammalian Genomes" (New York 2004) where he presented his work. In 2003 he joined the group of Prof. Michael Zhang in Cold Spring Harbor Laboratory, New York, USA as a visiting scientist for a period of 3 months. During his PhD he acquired strong computational skills and employed databases such as 4D, MySQL and ZODB. He is a volunteer developer for the Biopython project and wrote the sequence utilities module (SeqUtil). Yair is currently pursuing a postdoc position in which he will develop bioinformatics tools to investigate and study diseases and biological systems.

ACKNOWLEDGEMENTS

Since my early 20's I dreamed of going to study abroad. In the early days of my Pharmacy studies I was convinced I will go for a PhD in the United States, but life turns out to be what happens when you make other plans. Several coincidences lead me to come to the Netherlands and I would first like to acknowledge the Dutch people for their warmth, hospitality, peacefulness and their way of life that changed my perspective forever. I hold the greatest respect for them and this wonderful country that became my home for the past five years. I would like to thank my uncle Meyer for helping us make the transition period as smooth as possible and for taking care of us during this whole period. Although I ended up at Utrecht University I would like to thank from the bottom of my heart Hans Junginger, Ad Ijzerman and Douwe Breimer from Leiden University for the stimulating conversations, encouragement, kindness and support.

In April 2000 I met Ian for the first time. He immediately inspired the scientist in me and during the years I have learned much from him, from his superb art of writing for publications to his philosophy and ideas on how the universe works. I am eternally grateful for the opportunity of doing my PhD with him. In October 2000 I arrived to the lab at Utrecht University for the first time. Martin and Roel ruled the lab and guided me with patience and care. The Pharmaceutical Proteomics department quickly became my family. I greatly enjoyed the computer and gadget conversations with Rene; the lengthy discussion with Erik on everything from the Mac platform to Elispot assays; the amazement of discovering the Japanese culture through the eyes of Ryuji; and the stimulating conversations about tanks with Amos. The daily life in-front of the computer would not be same without David Gestel. He inspired me with his exquisite taste in music, supported and challenged me with his programming skills and was a great friend throughout the years. I am also grateful to Ed Moret and Monique Slijper for always being available for scientific discussions and their willingness to help whenever possible. I acknowledge Filip for sharing all the statistical courses with me and the occasional corridor conversations; and Esmail who warmed my heart with his kindness and friendship. Last and definitely not least, I would like to thank Ronald Oosting for being my mentor every step of my PhD period. I am grateful to him for believing in me, his dedication to my project until the very end and for guiding me through the complex world of molecular biology. I could not have done this without him.

In November 2003 I joined the group of Michael Zhang at Cold Spring Harbor Laboratory in New York as a visiting scientist. Michael, Zhenyu, Nila, Dustin, Pavel, Debo and Simon were like a family to me in this period of three months. They showed me how a bioinformatics group ticks and contributed immensely to my small genes project. This period was unforgettable and I am looking forward to future visits to CSHL. This trip to New York would not have been as easy and fun without Roushy and Andre Dahan and their children Maya and Michael. They embraced me with warmth and love and always made me feel at home.

I acknowledge BII (Biosystems Informatics Institute) for allowing me to access their Mac cluster which was essential for completing my work and especially Shane Sturrock for the constant patience and support.

I am grateful for our Dutch friends Petra & Leo, Inge & Peter and Rixt & Nico for all the fun times we had together and for being here for us. Ody and Anto, for their stunning internal energy, challenging political discussions and their friendship.

Although far from home, many of my friends in Israel kept in touch and were there for me throughout this period: Yaniv, Sagi, Ilan, Gal & Efrat, Rami & Revital, Noam & Ayelet, Shira & Giyora, Dana & Ronen, Guy & Romina, Shira & Yoav, Lior, Michal, Anat & Nir, Tali & Eran, Amiya & Moshe. I especially appreciate the computational input of Guy Meyer. He provided me with valuable feedback, contributed code and his suggestion were always right on target. He is my endless source of computational knowledge and I will always look up to him as the Mac master. I cherish my unique friendship with Galy & Zivi. They are as close as it comes to having soul mates and their affect on my life in the past years has been immense. I also thank Zivi for the wonderful design of this book's cover.

I am eternally grateful to have Maytal's family as my own. Dorit, who defies all the stereotypes of the mother in law. She is a second mother to me, a true friend who's always fun to be with, loving and caring. Avi, who I never had a chance to meet but lies in my heart and soul; Ben and Lihi for making me feel as their own flesh and blood; Shivi & Yankale for

their love, affection and the emotion they elicit in me every-time we get together; Yaara & Inbar for keeping our spirit alive with laughter and little patchuli for being my sweetest and most amazing “baby girl”; Shmulik for documenting every minute we missed; and Zvi and Ruth for their love and ability to make and nurture such an amazing family.

The Benita family, my family, is the solid ground I stand upon. My father, Shimon, who is a great source of inspiration and a role model. I cherish our unique father & son bond and am always amazed by his affection and support in my personal and professional life. My mother, Esther, for her sensitivity, love and emotional wisdom that shaped me as the man, son, husband and father I am today. My sister Einat for her wonderful sense of humor, love, care and sensitivity.

Maytal - you are the energy that keeps my blood flowing. Finding you was the single greatest accomplishment of my life. Our constant growing love knows no borders and every time I think nothing could be more perfect than this you show me that reality exceeds any imagination. As I am writing these words you are still recovering from the birth of our little baby, Ariel. You turned me into a father and us into a family. Once again, nothing could be more perfect.

A knowledge of the science of life is only made complete by a knowledge of those living it.

The human body is not just a biological machine,
it's much more than that-
it's a person

Robert Winston