ANALYSIS
# of the
HUMAN TRANSCRIPTOME
and
IDENTIFICATION
of
CONSERVED NONCODING ELEMENTS

by

Emiliano Dalla

PhD in FUNCTIONAL and
STRUCTURAL GENOMICS

International School for Advanced Studies
(ISAS/SISSA)

2005

Supervisor: Prof Claudio Schneider

External Supervisor: Prof Michael Q Zhang

TABLE OF CONTENTS

INTRODUCTION

The Human Genome Projects: the Public effort and the Celera consortium

The sequence of the human genome is of interest for different reasons: it is the largest genome to be extensively sequenced so far, being 25 times as large as any previously sequenced genome and eight times as large as the sum of all such genomes; it is the first vertebrate genome to be extensively sequenced and, uniquely, it is the genome of our own species. Along this, its analysis is going to allow to investigate the mechanisms underlying human evolution, the causation of diseases and the interplay between the environment and heredity in defining the human condition.

The beginning of year 2001 was characterized by the simultaneous publication of two different works concerning the initial sequencing and analysis of the human genome, one presented by the International Human Genome Sequencing Consortium (Lander ES et al, Nature 2001) and the other by Celera Genomics in collaboration with partners from both public and private institutes (Venter CJ et al, Science 2001).

One of the first feature the draft genome sequence makes it possible to explore is the variation in GC content, which appears to be extremely high, as there are huge regions (>10 Mb) with GC content far from the average. The so-called CpG islands (stretches of unmethylated DNA enriched in CpG dinucleotides) are particularly interesting because many are associated with the 5’ ends of genes (Bird AP, Trends Genet 1987), while others may be part of repeat elements (playing sometimes the role of control regions). The relative density of CpG islands correlates reasonably well with estimates of relative gene density on the corresponding chromosomes, based both on previous mapping studies involving ESTs (Fig.1) and on the distribution of gene predictions obtained by analyzing the draft genome sequence.

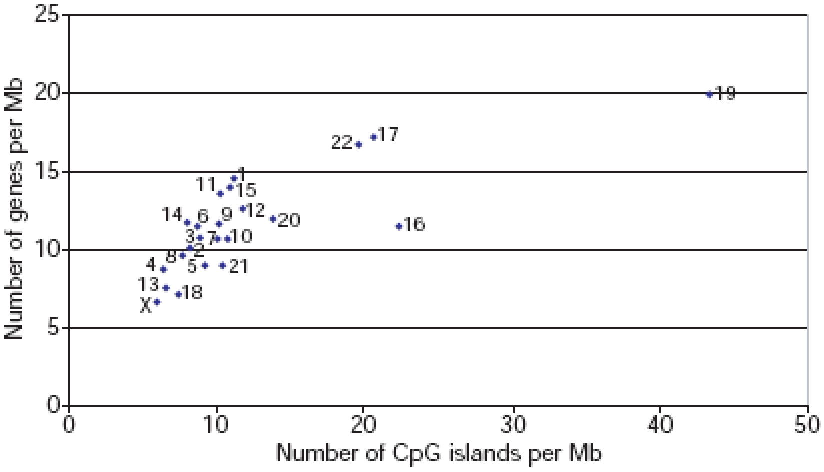


Figure 1. Number of CpG islands per Mb for each chromosome, plotted against the number of genes per Mb (From: *Lander ES et al, Nature 2001*).

On the basis of these first considerations, there seemed to be about 30,000-40,000 protein-coding genes in the human genome - only about twice as many as in worm or fly - corresponding to less than 5% of the human genome. However, more recent estimations that also introduced corrections for ploidy (Mattick JS, Nat Rev Genet 2004; Cheng J et al, Science 2005; Prasanth KV et al, Cell 2005) have lowered the number of protein-coding genes to 20,000-25,000 while increasing the overall gene number, so that actually it is thought that, although about 4% of the human/mouse/rat genomes are under selective pressure, only about 1.5% appear to be protein coding. Along that, the genes are more complex, with more alternative splicing events generating a larger number of protein products and with definitely more complex mechanisms of transcription regulation. Moreover, the identification of new human genes is complicated by the fact that they tend to have small exons (encoding an average of only 50 codons) separated by long introns (some exceeding 10 kb), which creates great accuracy problems to gene prediction algorithms. For this reason cDNA libraries are one of the more precious tools, as the comparison of cDNA sequences from different organisms can allow the attribution of functional roles to unknown human genome regions, although this method can only be applied to strongly conserved genes, while rapidly evolving genes are less identifiable.

The remaining ~98% of the human genome is made of non-coding sequences, more than 50% being represented by repeat sequences of different kinds. The distribution of these repeated elements at a small genome scale is strikingly variant: while some regions of the genome are extraordinarily dense, some other regions are nearly devoid of repeats. The absence of repeats may be a sign of large-scale *cis*-regulatory elements that cannot tolerate being interrupted by insertions. The four regions with the lowest density of interspersed repeats in the human genome are the four homeobox gene clusters: HOXA, HOXB, HOXC and HOXD. Each locus contains regions of around 100 kb containing less than 2% interspersed repeats. Ongoing sequence analysis of the four HOX clusters in different gnathostome lineages shows a similar absence of transposable elements, and reveals a high density of conserved noncoding elements (Chiu C et al, Proc Natl Acad Sci USA 2002). If this consideration is correct, exploring the human genome looking for similar regions may allow to identify new elements regulating human genes of major importance.

Noncoding transcribed regions

Recently there has been a blast in the discovery of transcribed molecules that play roles different from these of templates or members of the protein synthesis apparatus (Mattick JS, Nat Rev Genet 2004; Storz G et al, Annu Rev Biochem 2005; Griffiths-Jones S et al, Nucleic Acids Res 2005). Some of the best known members of this family of noncoding RNAs (ncRNAs) include small nucleolar RNAs (snoRNAs), necessary for molecular processing and modification in the nucleolus, and small nuclear RNAs (snRNAs), that play a critical role in the spliceosome activity. However, along with these critical components of cellular enzymes, a huge number of newly identified RNAs have been found to function as regulators in almost all the steps of gene expression pathways (Fig.2), like microRNAs (miRNAs) that are involved in the regulation of higher eukaryotes processes – such as development, cell-death and fat metabolism- by repressing the translation of their targets.

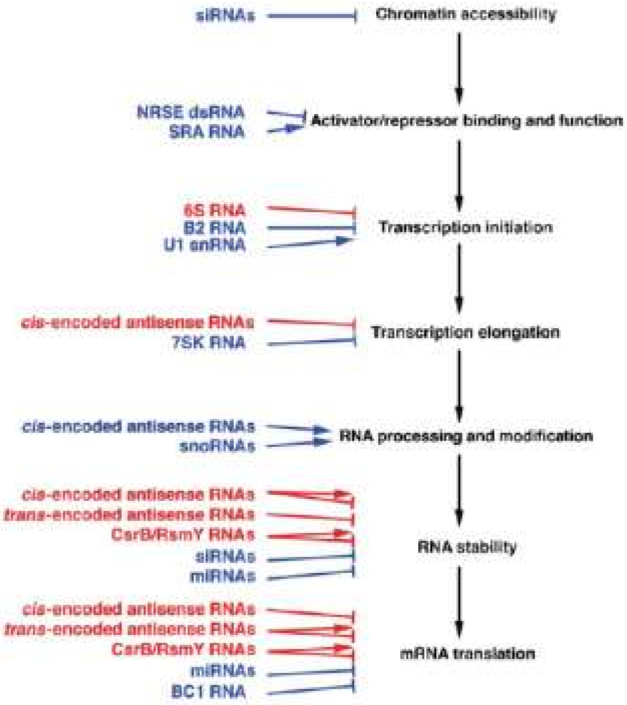

siRNAs
Chromatin accessibility
NRSE dsRNA
SRA RNA
Activator/repressor binding and function
6S RNA
B2 RNA
U1 snRNA
Transcription initiation
cis-encoded antisense RNAs
7SK RNA
Transcription elongation
cis-encoded antisense RNAs
snoRNAs
RNA processing and modification
cis-encoded antisense RNAs
trans-encoded antisense RNAs
CsrB/RsmY RNAs
siRNAs
miRNAs
RNA stability
cis-encoded antisense RNAs
trans-encoded antisense RNAs
CsrB/RsmY RNAs
miRNAs
BC1 RNA
mRNA translation

Figure 2. Steps in gene expression at which RNAs have been found to modulate gene expression. Bacterial and plasmid (*red*) and eukaryotic (*blue*) ncRNAs exert positive (*arrows*) and negative (*bars*) regulation at every step in gene expression (From: *Storz G et al, Annu Rev Biochem 2005*).

The existence of RNA regulatory molecules represents a great advantage, from an evolutionary point of view, as many of the ncRNAs known thus far are expressed and function during specific developmental stages or under stress conditions, when resources are limited and when the lower input of energy and shorter time required to synthesize a short ncRNA compared to a protein can prove critical for cell survival.

Actually, the best known mechanism by which ncRNAs fulfil their functional role is by base pairing with target transcripts. c*is*-encoded elements are encoded at the same genetic location, but on the opposite strand to the RNAs they act upon and therefore contain perfect complementarity to their target, while *trans*-encoded RNAs are encoded at a chromosomal location distinct from the RNAs they act upon and generally do not exhibit perfect base-pairing potential with their targets.

Among the human *cis*-encoded elements those that have received extensive attention are certainly the small interfering RNAs (siRNAs) (Montgomery MK, Methods Mol Biol 2004; Meister G and Tuschl T, Nature 2004).

These ~21–25nt RNA fragments are usually derived from double-stranded RNA of exogenous origin and are thought to be a defense against foreign RNA. However, there are examples in which sense and antisense transcripts derived from endogenous sequences (i.e. repeated sequences) can "silence" the expression of the sense transcript by targeting the sense transcript for degradation or by modulating the structure of the chromosome encoding the RNAs (Lippman Z and Martienssen R, Nature 2004). The details by which siRNAs modulate chromatin structure remain to be delineated, as well as only a very limited number of the other antisense transcripts have been shown to exert a clear regulatory effect on the *cis*-encoded mRNAs. One rare, known example is the Rev-ErbA*α* RNA, an antisense transcript detected in B lymphocytes. This antisense transcript overlaps one of two antagonistic sites in the mRNA encoding the thyroid hormone receptor, and expression of the Rev-ErbA*α* RNA correlates with a change in the ratio between the two splice forms of the thyroid hormone receptor transcript (Hastings ML et al, J Biol Chem 2000).

In contrast to the relatively few chromosomal *cis*-encoded antisense RNAs known to have function, many chromosomal *trans*-encoded antisense RNAs have been found to exert regulation in eukaryotes, although the reasons of this behaviour are not clear. The *trans*-encoded microRNAs (miRNAs) found in worms, flies, plants, and vertebrates base pair with mRNAs and modulate mRNA stability and translation, analogous to many of the bacterial *trans*-encoded antisense RNAs. The consequences of base pairing, although, remain to be understood, but in general it seems that the perfect or near-perfect base pairing associated with siRNAs and some miRNAs leads to RNA degradation, whereas the imperfect base pairing associated with other miRNAs leads to repression of translation. However, it is not known what other factors contribute to the outcome of ncRNA-target-RNA pairing, and another problem that deserves further attention is the question of how *trans*-encoded RNAs are turned over once the environmental or developmental signal leading to their induction is removed.

Along with ncRNAs that act by base pairing, it has been recently found that some other regulatory RNAs directly bind protein targets modifying their activity.

One possibility is that these RNAs bind and modulate the function of proteins involved in transcription, and different examples of this kind exist among human ncRNAs: the U1 snRNA, part of the spliceosome, also binds to the general transcription factor TFIIH stimulating transcription initiation in a way that is still unknown (Kwek KY et al, Nat Struct Biology 2002); the steroid receptor RNA activator (SRA) RNA acts as a co-activator to stimulate the transcription of steroid receptor-dependent genes, but its mechanism is unknown, too (Lanz RB et al, Cell 1999); finally, the NRSE dsRNA, a small, double-stranded RNA that is bound by the NRSF/REST protein (a negative transcriptional regulator that restricts neuronal gene expression to neurons) due to its similarity to the NRSF/REST-binding site and that also seems to be able to repress the activity of the NRSF/REST repressor, thus suggesting the hypothesis that NRSE dsRNA could compete for NRSF/REST binding to its promoter sequences (Kuwabara T et al, Cell 2004).

Another mechanism (although quite rare) is that some ncRNAs can bind and modify the activity of proteins that regulate mRNA stability and translation: the dendritic BC1 RNA appears to promote the interaction between FMRP and its target mRNAs, thus helping FMRP in inhibiting mRNA translation (Zalfa F et al, Cell 2003). This possibility to regulate gene expression at a post-transcriptional level is very important, as it ensures a fast response to developmental and environmental signals.

It also seems that during evolution, probably due to reverse transcription events involving ncRNA genes, a great number of pseudogenes appeared, some of which being able to regulate the expression of their homologous protein-coding genes (Hirotsune S et al, Nature 2003), and different kind of SINEs and Alu repeats seem to have been generated, too, in this way.

It is also important to remember that, thus far, the attention has always been focused mainly on polyadenylated RNAs that are processed and exported to the cytoplasm, but recently (Kampa D et al, Genome Res 2004) it has been demonstrated that the set of non-polyadenylated nuclear RNAs may be very large, and that many such transcripts arise from so-called intergenic regions (Cheng J et al, Science 2005), with 36.9% (called *bimorphic*) transcribed regions of the human genome existing also in the non-polyadenylated form and 43.7% (many with an exclusive nuclear location) existing ONLY in the latter status, suggesting that novel regulatory mechanisms may be involved in the identification of transcripts whose polyadenylation states are altered as means of regulation.

Finally, compartmentalization has been identified as another possible way of regulating gene expression. A ~8 kb nuclear-retained poly(A)$^+$ RNA has in fact been found (Prasanth KV et al, Cell 2005) to regulate the level of its protein-coding partner, the shifting from one transcript to the other one occurring by alternative promoter and poly(A) site usage. This work, then, suggests a new role of the nucleus in harbouring RNA molecules that are not immediately needed to produce proteins but whose cytoplasmic presence is rapidly required upon physiologic stress.

Until now, ncRNAs genes have been in general very hard to identify during genome sequence analysis owing to their lack of defined sequence features, and also because of their small size and because they are resistant to frameshift and nonsense mutations, which apply only to protein coding genes. For this reason, new methods are currently being developed for the detection of new ncRNAs, both computational (heavily based on sequence conservation, thus limited in their range of application) and direct (such as size fractionation of total RNA or direct cloning after size selection steps), although many aspects of ncRNAs has still to be unveiled before exhaustive prediction become possible.

Protein coding genes

Although it has been demonstrated that noncoding regions of the human genome contain elements that play fundamental roles in many processes (both structural and functional), protein coding regions are thus far those on which the attention has been mainly focused. The identification of protein-coding genes, however, is also one of the most difficult tasks.

The availability of full-length cDNA sequences, collected in the NCBI database called RefSeq (Pruitt KD et al, Nucleic Acids Res 2001), has allowed to draw the first, partial conclusions on the exon-intron structure of human genes. First of all, it seems that while there is considerable variation in overall gene size and intron size, the variation in the size distribution of coding sequences and exons is less extreme. This feature has also been verified in other organisms such as worm and fly, although in the latter cases the mean size for internal exons is larger (Fig.3a, b), thus suggesting the existence of an exon-driven component of the splicing machinery (Sterner DA et al, Proc Natl Acad Sci USA 1996).

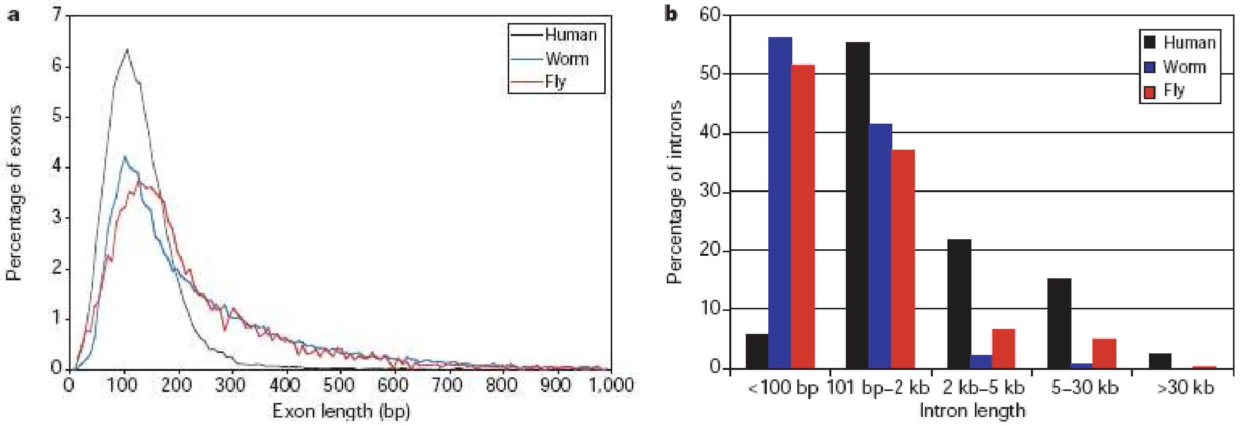


Figure 3. Size distribution of exons and introns in sequenced genomes (From: *Lander ES et al, Nature 2001*)**.**

RefSeq mRNAs mapping to the human draft genome sequence has also allowed to study the frequency of alternative splicing, an event that seem to have affirmed itself during evolution, reaching its top in mammals, and in particular in man, where it seems that at least 35% of genes undergo this process, compared to just 22% in worm. Accurate analysis has involved in particular human chromosomes 22 (whose complete sequence is available) and 19 (particularly gene-rich).

Chromosome 22 appears to have 60% of its genes involved in alternative splicing, with an average of 2.6 spliced transcripts per gene, while spliced genes located on chromosome 19 have an average of 3.2 splice forms. 70% of these splicing events occur inside the coding sequence, with 20% occurring in terminal exons (that in 24% of considered cases are also involved in alternative polyadenylation).

Unfortunately the RefSeq database contains only a fraction of all the existing human transcripts and is not exhaustive, *di per se*, in allowing to define the overall amount of human protein-coding genes. New gene identifying approaches, then, are required in order to derive more accurate estimates and in the past 10 years different types of data have been used, such as ESTs (Fields C et al, Nature Genet 1994; Liang F et al, Nature Genet 2000), cross-species genome comparisons (Roest Crollius H et al, Nature Genet 2000; Waterston RH et al, Nature 2002; Gibbs RA et al, Nature 2004) and *ab initio* analysis (Burge C and Karlin S, J Mol Biol 1997; Kulp D et al, ISMB 1996; Solovyev V and Salamov A, ISMB 1997) of finished chromosomes. Due to the extreme complexity of the human genome, these new approaches have not been able to completely solve the problem, as demonstrated by the work on human chromosomes 21 (Hattori M et al, Nature 2000) and 22 (Dunham I et al, Nature 1999), whose gene content still remain uncertain (with upper and lower estimates differing of as much as 30%) even with the availability of finished sequences and intensive experimental work. However, if conservative approaches are followed and only the most rigorous analyses are kept, all these methods tend to identify an overall number of 20,000-25,000 human genes, corresponding to 35,000-40,000 different proteins (Fig.4), with an average coding length of 1,400bp and average genomic extent of about 30kb

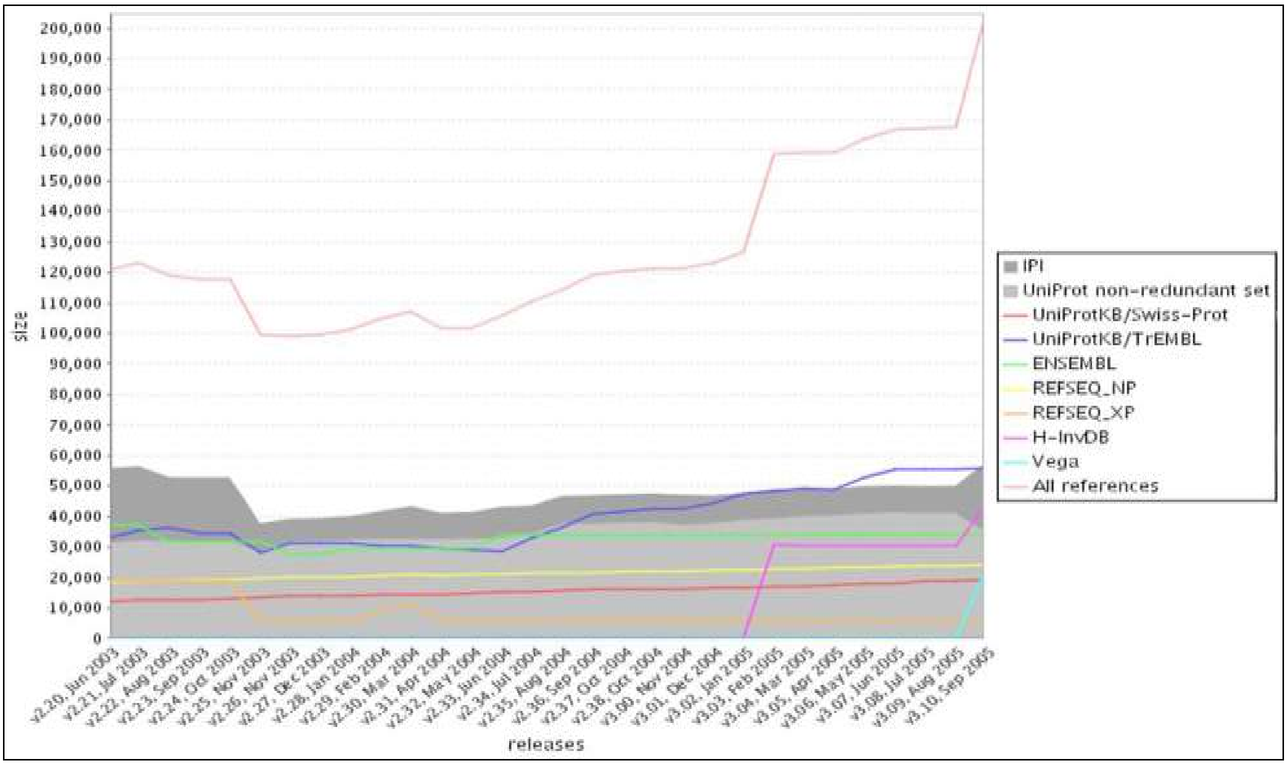


Figure 4. History of human IPI (v3.10, September 2005). IPI and UniProt sets (areas) and references source entries (lines) (From: *http://www.ebi.ac.uk/IPI/IPIhuman.html*).

It is important to remember that some classes of human genes may have been missed by all of the gene-finding methods. Genes could be missed if they are expressed at low levels or in rare tissues (being absent or very under-represented in EST and mRNA databases) and have sequences that evolve rapidly (being hard to detect by protein homology and genome comparison).

A particular case is represented by single-exon genes encoding small proteins, that may have been missed, too, because EST evidence that supports them cannot be distinguished from genomic contamination in the EST dataset and because homology may be hard to detect for small proteins (Basrai MA et al, Genome Res 1997). These genes, also called intronless paralogs (Elliott DJ et al, Hum Mol Genet 2000; Makeyev AV et al, J Biol Chem 1999), may be the result of retro-transposition events occurring to processed mRNA transcripts (the same mechanism is also involved in the production of pseudogenes) and seem to be able to encode the same proteins than the intron-containing forms of the same genes. Most of the single-exon genes identified so far are flanked by direct repeat sequences, although the precise nature of these repeats is not clear, and all of the cases for which there is high confidence contain poly(A) tails characteristic of retrotransposition. The distribution of intron-containing genes and of their intronless paralogs on human chromosomes is random, although there seem to be an over-representation of genes involved in translational processes and nuclear regulation, as well as metabolic and regulatory enzymes (Venter CJ et al, Science 2001). EST matches specific to a subset of intronless paralogs have been identified, suggesting the expression of these intronless paralogs, and differences in the upstream regulatory sequences between the source genes and their intronless paralogs could account for differences in tissue-specific gene expression. Defining which, if any, of these processed genes are functionally expressed and translated, however, will require further elucidation and experimental validation.

Finally, some genes may have been missed also because of the asymmetrical distribution of G+C content, CpG islands, and genes (Holmquist GP, Am J Hum Genet 1992). The genes, in fact, are not distributed quite as unequally as had been predicted: the most G+C-rich fraction of the genome constitutes more of the genome than previously thought (about 9%), and are the most gene-dense fraction, but contain only 25% of the genes, rather than the predicted ~40%. The low G+C, on the other hand, make up 65% of the genome, and 48% of the genes. This inhomogeneity, the net result of millions of years of mammalian gene duplication, has been described as the “desertification” of the vertebrate genome (Ohno S, Trends Genet 1985), and it seem that almost 20% of the human genome is in “deserts” (regions with >500 kbp without a gene). However, the apparent lack of predicted genes in these regions does not necessarily imply that they are devoid of biological function, as well as it is not clear if the “desertification” process occurred by accident or if it has been driven by selection and evolution.

Vertebrate Genome Projects: expanding the human genome knowledge

Although thousands of human genes have been identified and characterized during these years, over 40% of the predicted human protein set (including many products of disease genes of unknown biochemical function) cannot be ascribed a molecular function by methods that assign proteins to known families.

In order to fill the gaps that still remain in the knowledge of the human genome structure one powerful approach consists in producing and analysing cDNA libraries from multiple tissues of different organisms at suitable evolutionary distances to obtain a cross-species sequence comparison. Some of the best studied organisms include the laboratory mouse and the pufferfish, whose gene-specific patterns of alternative splicing are currently being studied along with human data.

Conservation between human and the pufferfish Tetraodon nigroviridis

One of the first organisms whose genome was studied with the aim of enriching the knowledge on the human genome was the pufferfish (Roest Crollius H et al, Nature Genet 2000), that diverged from human 400 Myr ago.

By exploiting the concept of "ecores" (evolutionary conserved regions) a mechanism called "Exofish" (exon finding by sequence homology) was applied to select those alignments between human and *T.negroviridis* DNA overlapping human exons. Chromosome 22, the first to be completely sequenced, was used as a starting point for the comparison and lead in 1999 to an estimation of 28,000 – 34,000 genes in the human genome. This finding, thus far unexpected, suggested that organism complexity is not a direct consequence of gene number, but has its source in other mechanisms that may include alternative splicing and multi-domain proteins. Almost one-third of human genes could not be detected at the time, including those for which the corresponding *T. nigroviridis* sequence was not yet known, those that evolve rapidly and for which protein sequence similarity is weak, and those that are strictly specific to mammals. However, the compactness of the *T. nigroviridis* genome was very useful to confirm that several neighbouring ecores that fell outside of existing annotations did belong to the same gene, as confirmed afterward by the cloning of new human cDNAs. By contrast, ecores identified inside the boundaries of the 545 annotated genes, but outside exons (that is, in introns), would correspond to exons that remained undetected by other homology-based approaches, presumably because of alternative splicing, as lately verified by other genome projects.

Conservation between human and mouse

The second complete mammalian genome to be revealed after human was the mouse.

The evolutionary distance between human and mouse has been evaluated in ~75 Myr. However, a significant fraction (over 90%) of the genomes shown synteny between the two (Fig.5), being preserved within conserved segments (Waterston RH et al, Nature 2002).

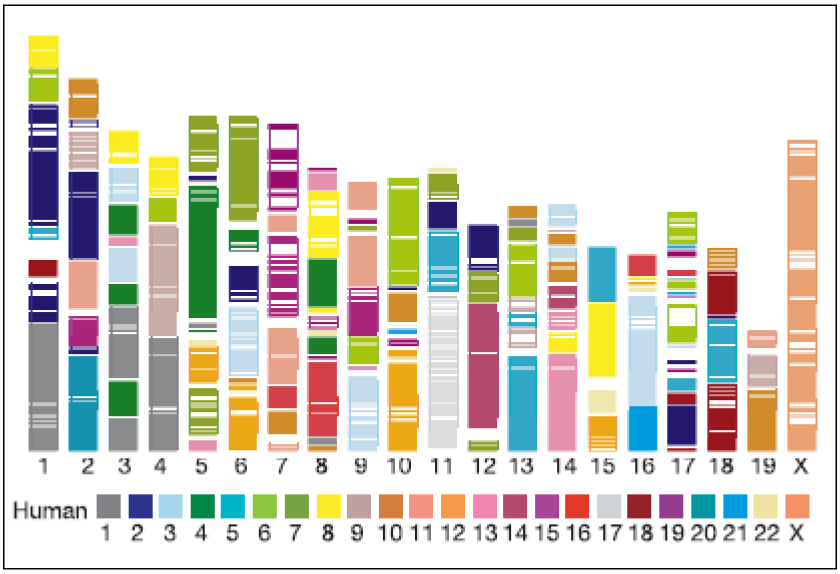

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X
Human
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X

Figure 5. Segments with conserved synteny in human are superimposed on the mouse genome. Each colour corresponds to a particular human chromosome (From: *Waterston RH et al, Nature 2002*).

Studying conserved segments between human and mouse has several uses, such as exploiting the conservation of gene order to identify likely orthologues between the species (particularly when investigating disease phenotypes) or using the two genome sequences as complementary scaffolds to assembly detailed comparative maps. Last but not least, with knowledge of both genomes biomedical studies of human genes can be complemented by experimental manipulations of corresponding mouse genes to accelerate functional understanding. In this respect, the mouse is unsurpassed as a model system for probing mammalian biology and human disease (Rossant J and McKerlie C, Trends Mol Med 2001).

It seems that the proportion of mouse genes with a single identifiable orthologue in the human genome is approximately 80%, with less than 1% of genes being without any homologue currently detectable in the human genome (and vice versa), while another 5% of the mammalian genome is made of small segments (50-100 bp) under (purifying) selection. This proportion is much higher than can be explained by protein-coding sequences alone, implying that the genome contains many additional features (such as untranslated regions, regulatory elements, non-protein-coding genes, and chromosomal structural elements) under selection for biological function.

For this reason the attention has then been focused on regions located 200bp upstream of transcription start sites, where canonical promoter regions are located. Such analysis is necessarily limited by the fact that transcriptional start sites remain poorly defined for many genes, but by the way it has allowed to verify a significant degree of conservation, too, although not as significant as that observed for 3' UTR (where along with the polyadenylation signal other regulatory signals exist, such as those that affect mRNA stability and localization) and especially for 5' UTR (~ 75-76% conservation, higher than previously expected). It has also been shown that conservation levels vary regionally within the features of a 'typical' gene. Sequence identity rises gradually from a background level to 78% near the approximate transcription start site, where the level reaches a plateau. It is possible that sharper definitions of transcriptional start sites would allow the footprint of the TATA box and other common structures near the transcription start site to emerge. Conversely, many human promoters lack a TATA box, and transcription start at such promoters is not typically sharply defined (Suzuki Y et al, Nucleic Acids Res 2004).

By increasing the number of regulatory control regions taken into account (Wasserman WW et al, Nature Genet 2000; Loots GG et al, Genome Res 2002), and in particular by enriching the contribution of elements < 2kb far from promoters, the extent of conservation increases, reaching levels similar to 5' UTRs.

Finally, to have a global overview of the "functional" conserved subset of the mammalian genome as a whole, an estimation has been made on the proportion of the genome that is better conserved than would be expected given the underlying neutral rate of substitution. The analysis thus suggests that about 5% of small segments (50 bp) in the human genome are under evolutionary selection for biological functions common to human and mouse, less than half corresponding to coding sequences. This corresponds to regions totalling about 140Mb of human genomic DNA, although not all of the nucleotides in these windows are under selection. In addition, some bases outside these windows are likely to be under selection, too. Conserved, non-coding sequences may for instance be sequences that control gene expression, such as the control element for the IGFALS gene (Fig.6); if a typical gene contains a few such regulatory sequences, there may be tens to hundreds of thousands of such elements interspersed in the whole human genome (Dermitzakis ET et al, Nature 2002). Furthermore, some of the conserved fraction may correspond to sequences that were under selection for some period of time but are no longer functional.

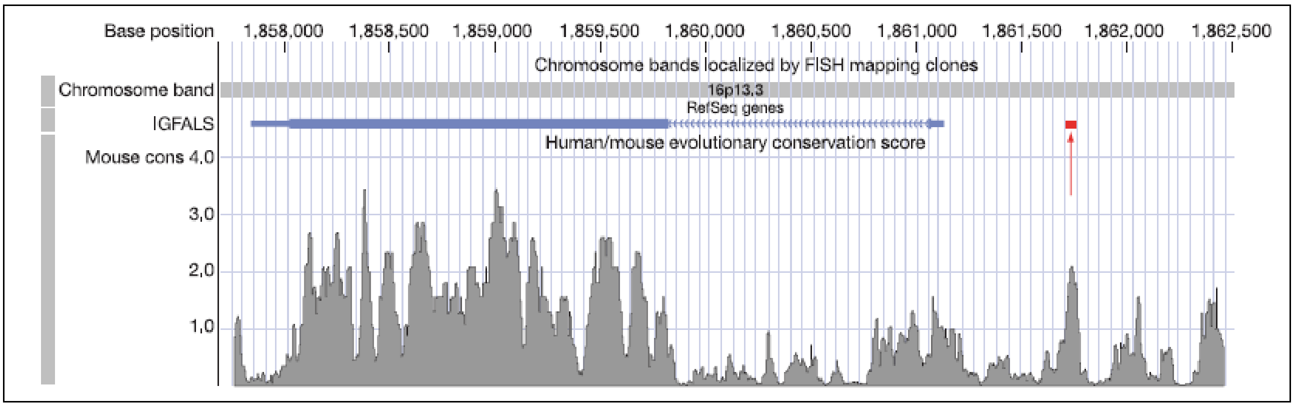


Figure 6. Conservation scores for 50-bp windows in a 4.5-kb region containing the human insulin-like growth factor binding protein acid labile subunit (IGFALS) gene. The two coding exons are displayed as taller blue rectangles, UTRs as shorter rectangles, and the intron is shown as a barbed line indicating direction of transcription. The red bar shows the location of the interferon-g-activated sequence-like element (GLE), which is bound by transcription factors from the STAT5a and STAT5b protein family to control expression of this gene. Additional regulatory elements may be located in the other peaks of conservation (From: *Waterston RH et al, Nature 2002*).

Of course, it should also be noted that non-conserved sequence may have important roles, too, for example, as a passive spacer or providing a function specific to one lineage.

It should certainly be possible to pinpoint these regulatory elements more precisely with the availability of additional related genomes. However, mouse is likely to provide the most powerful experimental platform for generating and testing hypotheses about their function. An example is the demonstration, based on mouse–human sequence alignment followed by knockout manipulation, of several long-range locus control regions that affect expression of the Il4/Il13/Il5 cluster (Loots GG et al, Science 2000).

Conservation between human and rat

The Brown Norway rat was the third complete mammalian genome to be deciphered, and three-way comparisons with the human and mouse genomes allowed to resolve more details of mammalian evolution.

First of all, this comparison confirmed that, despite the slight different size of the three genomes, they encode similar numbers of genes (Gibbs RA et al, Nature 2004), the number of coding exons per gene and average exon length being similar in the three species, and no significant deletion or duplication having occurred since their divergence. A billion nucleotides (~ 40% of the euchromatic rat genome) align orthologously to mouse and human and contains the vast majority (94-95%) of exons and known regulatory elements (1–2% of the genome). Another feature that was evaluated was the level of three-way conservation observed between the human, mouse and rat genomes in the ancestral core (Fig.7), to estimate the fraction of the human genome that is accumulating substitutions more slowly than the neutral rate in the three lineages since their divergence, and hence may be under some level of purifying selection. The results confirmed what had been previously predicted by a human-mouse comparison, giving values in the range of 5–6% when measured by two quite different methods (Cooper GM et al, Genome Res 2004). In this constrained fraction, non-coding regions outnumber coding regions regardless of the strength of constraint, an observation that supports recent comparative analyses limited to subsets of the genome. The preponderance of non-coding elements in the most constrained fraction of the genome underscores the likelihood that they play critical roles in mammalian biology (Dermitzakis ET et al, Nature 2002).

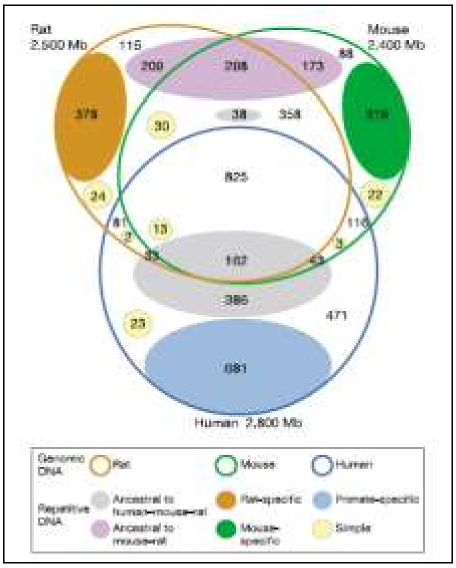


Figure 7. Aligning portions and origins of sequences in rat, mouse and human genomes. Uncoloured areas are non-repetitive DNA — the bulk is assumed to be ancestral to the human–rodent divergence. Numbers of nucleotides (in Mb) are given for each sector (From: *Gibbs RA et al, Nature 2004*).

Being the third mammal to be fully sequenced, the rat can also add significantly to the utility of nucleotide alignments for identifying conserved non-coding sequences (Tagle DA et al, J Mol Biol 1988; Gumuccio DL et al, Mol Cell Biol 1992; Boffelli D et al, Science 2003). This power increases roughly as a function of the total amount of neutral substitution represented in the alignment, and rat adds about 15% to the human–mouse comparison. Many conserved mammalian non-coding sequences are expected to have regulatory function, and can be predicted using further analyses based upon these alignments. These methods were thus applied for detecting significantly conserved elements and scoring regulatory potential (Kolbe D et al, Genome Res 2004) to the genome-wide human–mouse–rat alignments. Requiring conservation among mammalian genomes, in fact, greatly increases the specificity of predictions of transcription factor binding sites. Transcription factor databases such as TRANSFAC (Wingender E et al, Nucleic Acid Res 2001) contain known transcription factor binding sites and some knowledge of their distribution, but simply searching a sequence with these motifs provides little discriminatory power. For example, it has been demonstrated that all known regulatory elements and functional promoters have TRANSFAC matches, but so do 99% of the 2,049,195 mammalian ancestral repeats, most representing false-positive predictions. The introduction of conservation as a criterion for regulatory element identification greatly increases specificity, with only a modest cost in sensitivity. By demanding that the TRANSFAC matches are present and orthologously aligned in all three species— human, mouse and rat— then only 268 matches are recorded in ancestral repeats (0.01%), while 63 (74%) of the matches in the aforementioned known regulatory elements and 121 (80%) in functional promoters are retained. In general it has been shown that demanding conservation in the human–mouse–rat three-way alignments lead to a 44-fold increase in specificity.

Typical results show strong conservation for a coding exon, as well as for several non-coding regions (Fig.8).

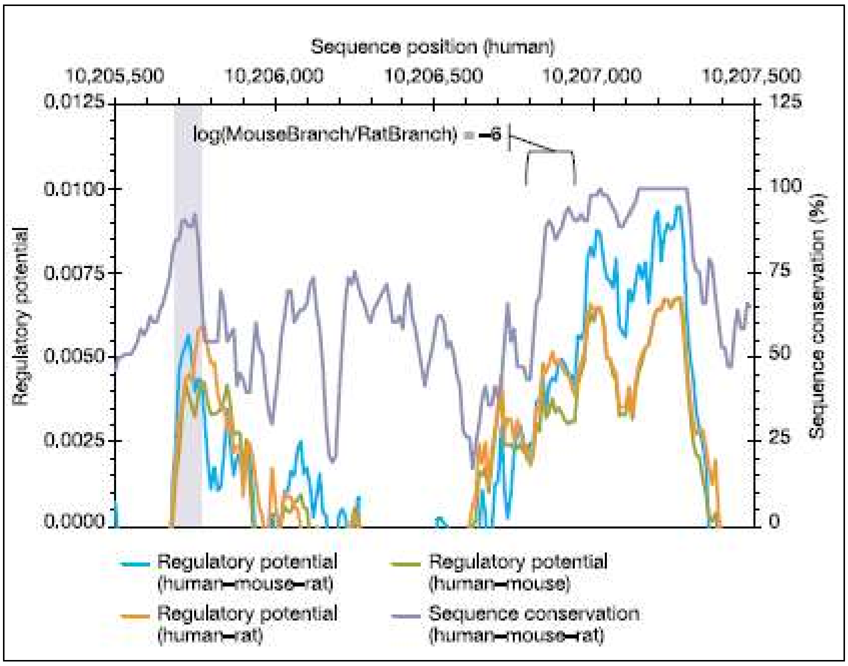


Figure 8. Close-up of PEX14 (peroxisomal membrane protein) locus on human chromosome 1 (with homologous mouse chromosome 4 and rat chromosome 5). Conservation score computed on three-way human–mouse–rat alignments presents a clear coding exon peak (grey bar) and very high values in a 504 bp non-coding, intronic segment (right; last 100 bp of alignment are identical in all three organisms) (From: *Gibbs RA et al, Nature 2004*).

Mammalian Transcriptome Projects: depicting the full coding potential… and more!

In the past years, along with projects aiming at studying the overall structure and composition of genomic DNA, many others started in order to analyse in more detail the transcribed fraction. These projects, initially launched to identify the protein-coding regions interspersed among the non-coding material, quickly became invaluable tools for studying the complex mechanisms occurring during gene expression. All these approaches were characterized by the fact of using cDNA libraries, collections of DNA molecules complementary to the mRNAs expressed in a given tissue (or cell-line) of a given organism at a given developmental stage. The production and sequencing of cDNA collections were obtained using different approaches but always converged to the same final result made of lists of possibly full-length Open Reading Frames (ORFs) that had to be mapped to the corresponding genes on the corresponding genome sequence. After a first phase made of random cloning, the attention is now being focused on protocols allowing to target the missing transcripts.

The Mammalian Gene Collection

This project, born to generate and sequence a publicly accessible cDNA resource containing a complete open reading frame (ORF) for every human and mouse gene, has actually led to the isolation of more than 11,000 human and 10,000 mouse genes represented by at least one clone with full ORF, whose coding potential has also been tested (Gerhard DS et al, Genome Res 2004). Recently, a rat cDNA component was added to the project, and ongoing frog (*Xenopus*) and zebrafish (*Danio*) cDNA projects were expanded to take advantage of the MGC experience. The random transcript sampling approach used thus far led to an under-representation of large and/or rare transcripts, as expected, and new, more directed cloning approaches are being used to solve this problem (50-80% of missing genes seem to be recoverable by using gene-specific primers (Baross A et al, Genome Res 2004)). In addition, a large number of previously uncharacterized full-ORF sequences, whose function is still unknown, have also been recovered and are going to be studied along with two other families of missing genes, the putative, computer-predicted genes for which there is some experimental evidence for the transcript's existence (such as one or more ESTs, or an uncharacterized cDNA generated through a large-scale project) and the ab-initio predicted genes based solely on computational methods. No information is currently available about non-coding transcripts and regulatory elements.

The FLJ Collection

The FLJ (Full-length Long Japan) collection started almost in the meantime than the MGC project and shares the same goals. The first characterization of the collected cDNAs (Ota T et al, Nat Genet 2004) led to the selection and sequencing of 21,243 clones corresponding to 10,897 non-redundant clusters with no evident match to known transcripts. Half of them seemed to be protein-coding, including 1,999 clusters that had never been predicted by computational methods and that presented a GC content ~58%, suggesting the existence of a slight bias against GC-rich transcripts in current gene prediction algorithms. The remaining 5,481 cDNAs contained no obvious open reading frames and were classified as putative ncRNAs, also due to the presence of clear patterns of splicing in ~25% of them and because of positive results from RT-PCR experiments; their GC content appeared to be different from that of new protein-coding cDNAs, suggesting the involvement of different region of the human genome, and their mapping regions were usually located more than 5 kb upstream of any known or predicted genes. As in the case of the MGC project, FLJ lacks cDNAs derived from long and/or rare mRNAs (i.e. those originated from small organs or rare cell types) and technical development will be needed to solve these problems.

The RIKEN Mouse Gene Encyclopaedia Project

The aim of this project, also known with the name “FANTOM” (Functional ANnoTation Of Mouse), can be considered the same as that of the MGC and FLJ: determining the full coding potential of the mouse genome by collecting and sequencing full-length complementary DNAs and physically mapping the corresponding genes to the mouse genome. However, one important aspect differentiates FANTOM from any previous similar project: given the limits of the semi-automated approach used to analyze the mouse cDNAs, it was thought that the best results would have been obtained if that automated analyses was followed by manual curation performed by experts in the fields of bioinformatics, genome science and biology. An international functional annotation meeting was then organized and 21,076 previously selected cDNAs were analyzed by using a web-based annotation interface (Kawai J et al, Nature 2001); this first round of analysis clearly validated the overall strategy but also suggested the need for further refinements (i.e. at least some unclassifiable transcripts probably represented unprocessed nuclear RNA that could have been potentially avoided by isolating cytoplasmic RNA).

For this reason one year later the second phase of the project (FANTOM2) started by re-analyzing all the 21,076 cDNAs that had already been previously studied, and by adding 39,694 new molecules, raising the pool of examined full-length cDNAs to 60,770 (Okazaki Y et al, Nature 2002; see also *Personal Publications*). These transcripts were clustered into 33,409 “transcriptional units" (a computational definition indicating a cluster of transcripts that contains a common core of genetic information, in some cases corresponding to a protein-coding region), among which 4,258 were new protein-coding messages.

This project, actually representing the most comprehensive survey of a mammalian transcriptome ever made, recently underwent its third phase during which the number of independent analysed transcripts was raised to 181,047 (Carninci P et al, Science 2005; see also *Personal Publications*), about ten times more than the estimated gene number in mouse and human. As expected, 16,247 new mouse protein-coding transcripts were isolated, including 5154 encoding previously unidentified proteins, many transcripts arising from mechanisms such as alternative promoter usage, splicing (whose frequency increased from 41% to 65%, with respect to the FANTOM2 analysis, and which seems to regulate in many cases protein domains content and organization, as among the protein-coding transcripts 79% of these splice variations alter the protein product) and polyadenylation (27 motif families with a putative modulator activity on polyadenylation were identified).

Confirming what had already been observed in Drosophila (Spellman PT and Rubin GM, J Biol 2002), the genomic mapping of the mouse transcriptome revealed the existence of "transcriptional forests", with overlapping transcription on both strands, without gaps, separated by "deserts" in which few transcripts were found. Thanks to CAGE technology (Shiraki T et al, Proc Natl Acad Sci USA 2003) it was also possible to identify the 3' UTRs of protein coding loci as the initiation start sites for the transcription of many ncRNAs, thus allowing to classify as ncRNAs many transcripts that had been previously considered truncated coding mRNAs. This was an outstanding discovery, as non-protein-coding RNAs represent one of the more significant classes of "genes" missing from the existing genome annotation. These RNAs, along with their involvement in protein synthesis (rRNAs and tRNAs), are also implicated in control processes such as genomic imprinting and perhaps more globally in control of genetic networks (Mattick JS and Gagen MJ, Mol Biol Evol 2001).

Globally, FANTOM3 led to the classification of 34,030 transcripts as ncRNAs (many of which matching to rat, mouse and/or human ESTs, presenting CpG islands in 5' upstream regions, being subject to splicing process and having an antisense location); their function has not been elucidated, yet, although it was possible to demonstrate their positional conservation across species. Moreover, their promoters were highly conserved, too, even more than those of protein-coding mRNAs (Fig.10a, b), and they contain binding sites for known transcription factors (Cawley S et al, Cell 2004). Their transcription may either be important for or be a consequence of genomic structure or sequence, or the transcript may act through some kind of sequence-specific interaction with the DNA sequence from which it is derived or with other targets that have not been identified, yet.

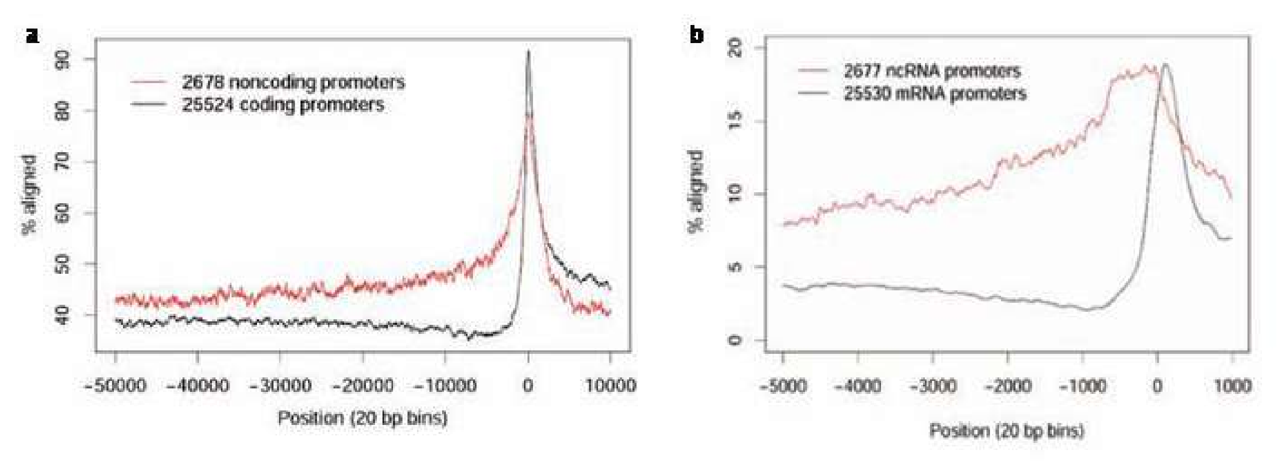
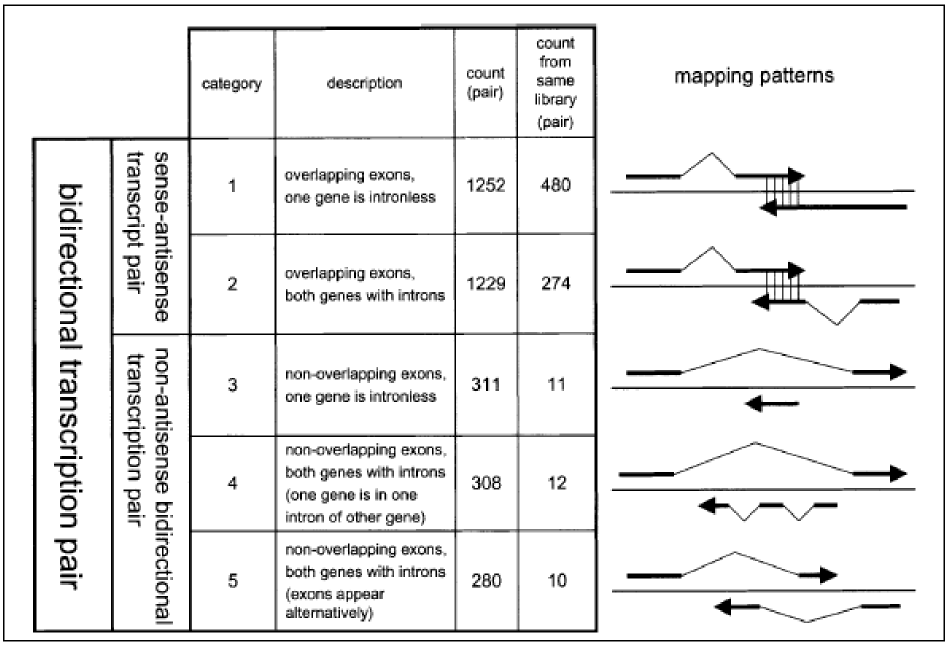


Figure 10. Noncoding RNA promoters conservation evaluated by alignment. a: mouse-human comparison. **b**: mouse-chicken comparison (From: *Carninci P et al, Science 2005*)**.**

During FANTOM3 there was also a re-evaluation of the sense-antisense (S/AS) pair phenomenon (Katayama S et al, Science 2005), with evidences suggesting it to be more widespread than previously thought: 4520 transcriptional units (TUs) were found to contain full-length transcripts forming S/AS pair on exons, whereas 4129 more TUs were transcribed from different strands of the same locus without apparently sharing overlapping exons, the first being characterized by a significantly lower mapping rate on chromosome X than the latter. Besides, sense and antisense transcripts tended to be isolated from the same libraries, while non-antisense bidirectional transcription pairs were not apparently co-regulated, thus suggesting a different basic, biological nature of the two types of transcription.

By considering these results, then, it seems clear that the existence of large numbers of natural antisense transcripts implies that the regulation of gene expression by antisense transcripts is more common that previously recognized, and in particular could alter transcription, elongation, processing, location stability, and translation. They may be coding or noncoding RNA (ncRNA) complementary to mature processed sense coding mRNA, or they may be complementary only to the primary unprocessed transcript, being contained solely within an intron or overlapping a 5’ UTR or 3’ UTR (Fig.9). They may also or may not be spliced.



		category	description	count (pair)	count from same library (pair)	mapping patterns
bidirectional transcription pair	sense-antisense transcript pair	1	overlapping exons, one gene is intronless	1252	480	
		2	overlapping exons, both genes with introns	1229	274	
	non-antisense bidirectional transcription pair	3	non-overlapping exons, one gene is intronless	311	11	
		4	non-overlapping exons, both genes with introns (one gene is in one intron of other gene)	308	12	
		5	non-overlapping exons, both genes with introns (exons appear alternatively)	280	10	

Figure 9. Classification of sense–antisense transcript and non-antisense bidirectional transcription pair patterns (From: *Kiyosawa H et al, Genome Research 2003*)**.**

Despite the mechanism, there are now many examples of functional antisense transcripts in developmental gene regulation and imprinting, although the number of well-characterized antisense transcripts is still small.

Although conservative, the combined S/AS prediction is 1.5- to 2-fold greater than that expected at the end of FANTOM2; moreover, S/AS interaction might also occur between immature RNAs (hnRNAs) in the nucleus, or introns themselves could originate smaller RNA with biological activity. The S/AS pair distribution didn’t seem to be random, as some chromosome showed greater or lower than average pair density, as well as their tissue regulation is not, as there seemed to be an over-representation for cytoplasmic proteins and under-representation for membrane and extracellular proteins.

Gene expression regulation: transcription factors and transcription factor binding sites

The one-dimensional script of the human genome, shared by essentially all cells in all tissues, contains sufficient information to provide for differentiation of hundreds of different cell types, and the ability to respond to a vast array of internal and external influences. Much of this plasticity results from the carefully orchestrated symphony of transcriptional regulation made of a network of interactions between proteins, called transcription factors (TFs), and their targets on DNA, called transcription factor binding sites (TFBS).

Much is known about the general process of DNA transcription and about the changes occurring in gene expression with respect to changes in the environment or to developmental constraints, but what is not clear, yet, is the mechanism that, given a quite limited battery of TFs, leads to the fine regulation of the expression levels of a complex genome like the mammalian genome. Along this, however, it is necessary to remember that although much has been learned about the *cis*-acting regulatory motifs of some specific genes, the regulatory signals for most genes still remain uncharacterized.

The core of the transcriptional apparatus (Fig.11) in mammals is made by RNA Polymerase II, the enzyme directly responsible of the synthesis of RNA from a DNA template, and by a small number of so-called *general transcription factors*, which must be assembled at the promoter before transcription can begin. This assembly process provides, in principle, multiple steps at which the rate of transcription initiation can be speeded up or slowed down in response to regulatory signals, and many eukaryotic gene regulatory proteins influence these steps.

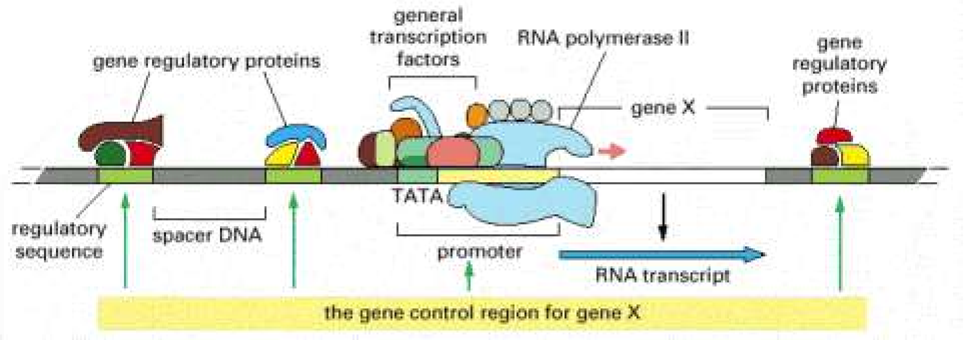


Figure 11. The gene control region of a typical eukaryotic gene. The gene control region of a typical eukaryotic gene. The *promoter* is the DNA sequence where the general transcription factors and the polymerase assemble. The *regulatory sequences* serve as binding sites for gene regulatory proteins, whose presence on the DNA affects the rate of transcription initiation (From: *Alberts B et al, Molecular Biology of the Cell, 4th Edition, Garland Publishing 2002*).

Transcription factors, more generally, bind *regulatory sequences* whose presence on the DNA affects the rate of transcription initiation. These sequences can be located adjacent to the promoter, far upstream of it, or even within introns or downstream of the gene, and the fact that transcription factors can act even when they are bound to DNA thousands of nucleotide pairs away from the promoter that they influence means that a single promoter can be controlled by an almost unlimited number of regulatory sequences scattered along the DNA, as the mechanism of DNA looping allow gene regulatory proteins bound at any of these positions to interact with the proteins that assemble at the promoter. Whereas the general transcription factors that assemble at the promoter are similar for all polymerase II transcribed genes, the gene regulatory proteins and the locations of their binding sites relative to the promoter are different for each gene. Many gene regulatory proteins, for instance, bind to enhancer sequences and activate gene transcription, while many others function as negative regulators (Fig.12); among the roughly 20,000-25,000 human genes, an estimated 5–10% encode gene regulatory proteins. These regulatory proteins vary from one gene control region to the next, and each is usually present in very small amounts in a cell, often less than 0.01% of the total protein. Most of them recognize their specific DNA sequences using one of their DNA-binding motifs but sometimes some of them do not recognize DNA directly but instead assemble on other DNA-bound proteins.

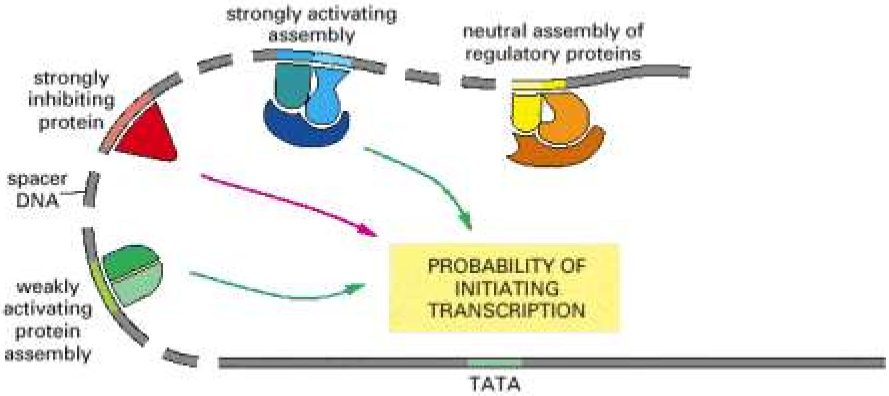

strongly activating assembly
neutral assembly of regulatory proteins
strongly inhibiting protein
spacer DNA
weakly activating protein assembly
PROBABILITY OF INITIATING TRANSCRIPTION
TATA

Figure 12. Integration at a promoter. Multiple sets of gene regulatory proteins can work together to influence transcription initiation at a promoter. It is not yet understood in detail how the integration of multiple inputs is achieved, but it is likely that the final transcriptional activity of the gene results from a competition between activators and repressors (From: *Alberts B et al, Molecular Biology of the Cell, 4th Edition, Garland Publishing 2002*).

Some of the best known gene activator proteins are those involved in histone acetylation and nucleosome remodelling, events that generally render the DNA packaged in chromatin more accessible to other proteins in the cell, including those required for transcription initiation. In addition, specific patterns of histone modification directly aid in the assembly of the general transcription factors at the promoter. Transcription initiation and the formation of a compact chromatin structure can be regarded as competing biochemical assembly reactions, so enzymes that increase, even transiently, the accessibility of DNA in chromatin will tend to assist transcription initiation. Gene activator proteins can sometimes act together to enhance a reaction rate (*transcriptional synergy*) and the joint effect is generally not merely the sum of the enhancements caused by each factor alone, but the product, or at least much higher than that produced by any of the activators working alone. Transcriptional synergy is observed both between different gene activator proteins bound upstream of a gene and between multiple DNA-bound molecules of the same activator. It is therefore not difficult to see how multiple gene regulatory proteins, each binding to a different regulatory DNA sequence, could control the final rate of transcription of a eukaryotic gene. For what concerns gene repressor proteins, the number of available mechanisms is as variable as in the case of activators, and some of them include the conversion of whole regions of chromosomes into heterochromatin, a form of chromatin that is normally resistant to transcription, or more fine regulations like those that lead to local repression of transcription of just a few genes. Like gene activator proteins, many eukaryotic repressor proteins act through more than one mechanism, thereby ensuring robust and efficient repression. A third category is made of gene regulatory proteins that do not themselves bind DNA but assemble on DNA-bound gene regulatory proteins: they are often termed co-activators or co-repressors, depending on their effect on transcription initiation, and as DNA-binding regulators can typically carry out multiple functions.

Typically, the assembly of a group of regulatory proteins on DNA is guided by a few relatively short stretches of nucleotide sequence. However, in some cases, a more elaborate protein-DNA structure, termed an *enhancesome*, is formed. A hallmark of enhancesomes is the participation of architectural proteins that bend the DNA by a defined angle and thereby promote the assembly of the other enhancesome proteins. Since formation of the enhancesome requires the presence of many gene regulatory proteins, it provides a simple way to ensure that a gene is expressed only when the correct combination of these proteins is present in the cell.

Different selections of gene regulatory proteins are present in different cell types and thereby direct the patterns of gene expression that give each cell type its unique characteristics. Each gene in a eukaryotic cell is regulated differently from nearly every other gene. Given the number of genes in eukaryotes and the complexity of their regulation, it has been difficult to formulate simple rules for gene regulation that apply in every case. For the same reason, it is also very difficult to formulate simple rules allowing to automatically recover, with the help of bioinformatics tools, the regulatory sequences involved in the expression of a given phenotype.

Along with its distance from the genes it regulates, another important feature of a promoter is its orientation, as this determines the orientation that RNA pol II will assume during transcription and, as a consequence, which of the two DNA strands is to serve as a template for the synthesis of RNA.

In general, it is possible to say that along with the TFBS sequences *di per se*, three features that play an important role in transcription regulation are TFBS structure, distance and order.

Thanks to the increasing number of DNA crystal structures determined it was possible to recognize general patterns of sequence specific helix distortion (Dickerson RE, Meth Enzymol 1992; Lu XJ and Olson WK, J Mol Biol 1999; Allemann RK and Egli M, Chem Biol 1997). Electrophoresis studies also demonstrated sequence dependent helix variations in DNA fragment mobility attribute to helix binding, thus it seems that groove width of helix might therefore be an important characteristic used by proteins for the recognition of specific sequences; due to the fact that TFBS are not always well conserved in primary sequence, conformational variations may play a role, too, in TFBS recognition and the use of this structure information could improve the predictive power of TFBS models. This was actually demonstrated in the case of the MetJ transcription factor of E.coli (Liu et al, Bioinformatics 2001), where the conformational model appeared to recognize features of TFBS different from those recognized by primary sequence based profiles. Combining the two approaches should then yield a hybrid model with improved discriminatory power compared to the single method alone.

In addition, it is known that specific arrangements of binding motifs within the regulatory regions (regulatory clusters) are necessary to achieve proper biological function. For example, an optimal spacing between binding sites (NF-Y and SRE motifs) was demonstrated in the human SREBP-2 promoter (sterol regulatory element-binding protein; Inoue J et al, J Biochem 1998). In vitro analysis of binding site arrangements in the rat collagenase-3 promoter (D'Alonzo RC et al, J Biol Chem 2002) revealed that a 10 bp (`helical') phasing in binding site distribution provides maximal transcriptional activity. The importance of the `helical phasing' and specific binding site arrangement was also demonstrated in vivo for the murine CD4 promoter (Sarafova S and Siu G, Nucleic Acids Res, 2000). In some cases, even a very small difference in the distance between binding motifs results in dramatically different transcriptional outcomes, one of the most striking examples of this kind being the binding of the POU domain transcription factor Pit-1 to its target sites, differentially spaced (2 bp difference) in growth hormone and prolactin gene promoters (Scully KM et al, Science 2000). In many of the described quantitative experimental studies, the disruption of specific spacing (phasing) between binding sites resulted in reduction, but not abolishment, of transcription, suggesting the presence of a certain flexibility in site arrangement. The biological reasons of this phenomenon are quite clear: the transcription factors, bound to promoter DNA, are also involved in specific protein-protein interactions; therefore, the binding motifs must be distributed in the promoter in a non-random fashion. In other words, the arrangement of binding motifs can control the formation of 3D protein complexes involved in initiation of specific transcription. Attempts to reveal and describe specific site arrangements resulted in a very interesting concept of composite elements (CEs) (Diamond MI et al, Science 1990). In the simplest case, a CE corresponds to a pair of individual binding motifs located at a particular distance and involved in formation of specific tertiary (DNA-protein-protein-DNA) complexes. Identical CEs may perform related functions in different genes. Currently, the CE concept is widely used for finding co-localized, synergistic (antagonistic) binding motif pairs or combinatorial arrays of motifs responsible for the formation of similar gene expression profiles. Another study conducted in Drosophila (Makeev VJ et al, Nucleic Acids Res 2003) explored preferential site distances in c*is*-regulatory modules (CRMs), transcription regulatory units (~1 kb range), often located far from the transcription start site and responsible for spatiotemporal expression of their cognate developmental genes, that resulted to have their binding sites arranged in particular ways, indicating the presence of specific developmental CEs. This results demonstrated that the binding sites were not distributed in a random fashion, with a large fraction of them belonging to small closely spaced groups (50-70 bp range). These findings confirmed the presence of different hierarchical levels of information, with CRMs constituting one of the upper levels, TFBS representing the bottom level and CEs being present in the “middle” of the hierarchy and acting independently in their response to the spectrum of native transcriptional signals. If this was true, repression of the same promoter (CRM) by two transcription factors might be achieved through an independent response of two or more corresponding CEs to the concentrations of these proteins, the maximal spatial independence of adjacent CEs being achieved through positioning of corresponding protein complexes on opposite sides of the DNA helix, and the existence of structurally different CEs (together with the different site affinity) might explain the differential gene response to concentrations of transcriptional regulators.

Finally, some evidence is starting to appear concerning the importance of the order of elements. By analysing different regions upstream of *S.cerevisiae* genes (Terai G and Takagi T, Bioinformatics 2004), several new element patterns that could not be found when considering a single kind of element or only combinations of elements were identified. For some of them the order was important without relation to the distance between them, for other instances the order and the closeness between elements were both important, and still others for which only the combination of elements or only the closeness were important, without relation to order. Accumulation of genome sequence data for closely related species will probably help in improving the accuracy if computational element prediction.

Transcription factor binding sites identification approaches

The identification of TFBS on the genome sequence is one of the most important tasks to perform in order to understand the mechanisms regulating the co-expression of a given set of genes.

The approaches that are currently used by promoter prediction programs (PPPs) to locate known and novel regulatory elements can be summarized in two categories:

i) comparative genomics (also called *phylogenetic footprinting*)

ii) single genomic approach

Comparative genomics

Studies of sequence alignment of regulatory domains of orthologous genes in multiple species have shown a remarkable correlation between sequence conservation, dubbed "phylogenetic footprints" (Tagle DA et al, J Mol Biol 1988), and the presence of binding motifs for transcription factors (Fig.13).

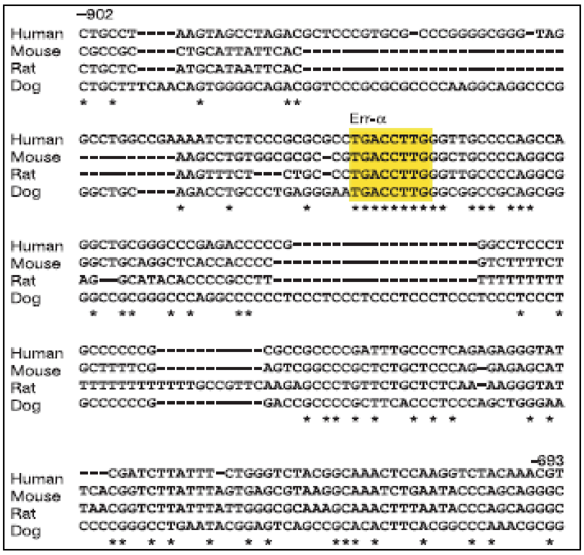


Figure 13. Conservation in GABPA promoter region reveals functional Err-a motif. Asterisks denote conserved bases. The yellow box marks the experimentally validated Err-a-binding site (From: *Xie X et al, Nature 2005*).

This approach could be particularly powerful if combined with expression array technologies that identify cohorts of genes that are co-ordinately regulated, implicating a common set of *cis*-acting regulatory sequences. Unfortunately, the compared analysis of mammalian genomes showed on one hand that the degree of homology between the coding regions was quite high, but on the other hand the conservation dramatically fell in the non-coding regions (especially those quite far from transcription start site), where TFBS are mostly expected to be located. Along this, it is also necessary to remember, as previously stated, that the primary sequence of DNA is not sufficient alone to justify all the interactions occurring between a genome and all the proteins that, at different levels, are involved in the regulation of gene expression. However, this approach represents a useful way for identifying at least the more conserved regions that have been maintained during evolution and that could work as a starting point for more precise analysis.

Single genomic approach

The great majority of existing TFBS predictive tools are based on a single genomic approach.

The identification of TFBS on a single genome sequence is a very difficult task, as they are short signals (typically about 10 bp long) in the midst of a great amount of statistical noise (a typical input being one regulatory region of length 1-3,000 bp upstream of each gene). To make matters worse, there is sequence variability among the binding sites of a given TF, and the nature of the variability itself is not well understood.

Recognition technologies employed to fulfil this task make use of a great variety of approaches that can be based on neural networks (*Dragon PF*; Bajic VB et al, J Mol Graph Model 2003), linear and quadratic discriminant analysis (*FirstEF*; Davuluri RV at al, Nat Genet 2001), Relevance Vector Machine (*Eponine*; Down TA and Hubbard TJ, Genome Res 2002), statistical properties of promoter regions (*CpGProD*; Ponger L and Mouchiroud D, Bioinformatics 2002), interpolated Markov model (*McPromoter*; Ohler U et al, Genome Biol 2002) or a combination of these.

Despite the method used by every single tool, it is possible to say that almost all of them rely on previous knowledge, that can be represented by weight matrices built by exploiting the knowledge on known TFBS (each position of the putative TFBS is associated to scores indicating the probability of occurrence of each one of the four possible nucleotides) or by some form of training, like in the case of methods relying on neural networks.

Recently, two distinct assessments examined in an exhaustive way a great number of publicly available programs for TFBS search, giving very interesting results that will need to be considered by whoever decides to approach this field.

In the first analysis, performed by evaluating the level of complexity of the whole human genome (Bajic VB et al, Nat Biotechnol 2004), one of the major points was that none of the programs achieved simultaneously balanced sensitivity and positive predictive value, with the most serious prediction inaccuracy being represented by non-CpG-island-related promoters. It was also quite clear that, with the technology used during the test, it was not possible to extrapolate the performance of the programs to the whole human genome, probably because the structure of promoters on different chromosomes varies and this variation could not be covered by the algorithms used.

The second analysis (Tompa M et al, Nat Biotechnol 2005) represented a different way of approaching the problem of TFBS search. First of all, along with examining human sequences also three other eukaryotes were considered: yeast, fly and mouse. The regions in which TFBS were searched changed, too, as the whole genome approach was abandoned and only the upstream regions (between -3,000 bp and 0, with respect to transcription start site (TSS)) were considered. As in the first assessment the tools that were compared were based on different algorithmic principles, exploiting Gibbs sampling method (*AlignACE*; Hughes JD et al, J Mol Biol 2000), weight matrices (*ANN-Spec*; Workman CT and Stormo GD, Pac Symp Biocomput 2000), consensus-based method (*Weeder*; Pavesi G et al, Nucleic Acids Res 2004) or exhaustive search algorithm (*YMF*; Sinha S and Tompa M, Nucleic Acids Res 2003), and all of them did not use auxiliary information, such as comparative sequence analysis, mRNA expression levels or chromatin immunoprecipitation results. The dataset was made of real TFBS, extracted from TRANSFAC, that were planted at their own position and orientation in three different types of background sequence: a) the binding sites' real promoter sequences, b) randomly chosen promoter sequences from the same genome and c) sequences generated by a Markov chain of order 3. Six of them were from fly, 26 from human, 12 from mouse and 8 from yeast.

The results, obtained by selecting the single best motif for each data set, gave rise to very interesting considerations. To summarize, the absolute measures of correctness of these programs were low: site sensitivity was at most 0.22 and correlation coefficient was at most 0.20. Different explanations can be given, to justify what seems to be a very negative final result:

1) due to the very partial understanding of the biology underlying regulatory mechanisms, there is a lack of an absolute standard against which predictive tools can be measured
2) comparative sequence analysis, that in some cases has proven to be a very powerful predictive tool, could not be used to refine the results by a cross-species comparison
3) the source of the TFBS that were used, TRANSFAC, is far from being completely reliable, and its binding sites appear to be unusually long, indicating a possible lack of precision in the experimental methods used to define them

In general, as can be seen in Fig.14, it seems that the threshold of genomic complexity beyond which the accuracy of TFBS predicting tools dramatically falls is represented by the yeast. As a further demonstration of this assumption, the program that predicted with the highest accuracy most of the considered domains is based on oligo frequencies analysis of all the available upstream sequences of the same organism (yeast TFBS, in this case). However, many tools perform much better on the yeast data sets than on other species, suggesting that the metazoan genome complexity is far from being elucidated at a satisfying level. The final suggestion given by the authors, in a certain sense a further proof of the inadequacy of existing approaches, consist in removing the "real" dataset, made of existing TFBS planted into their real genomic neighbourhood, and using only the "virtual" dataset in order to test the accuracy of the predictions.

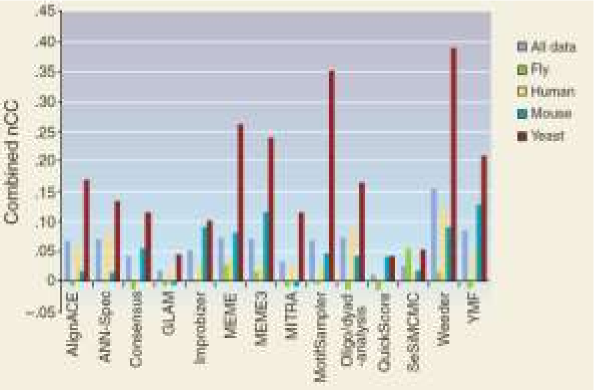


Figure 14. Accuracy comparison - Correlation coefficient by species (From: *Tompa M et al, Nat Biotechnol 2005*)**.**

The outcome of these two assessments clearly shows how the existing methods for TFBS prediction can't be used as models to develop new programs for identifying novel binding sites.

However, a recent work gives an even stronger confirmation of what was already clear, redefining the concept of promoter itself. With the aim of explaining into more details what had been previously observed (Kapranov P et al, Science 2002) during the mapping of transcribed regions on human chromosomes 21 and 22 (that is the existence of much more transcribed loci than could be accounted by the simple presence of protein coding RNAs), high-density tiled arrays representing essentially all non-repetitive sequences of these two chromosomes were used in a "ChIP-chip" experiment to map the binding sites of three major transcription factors: Sp1, cMyc and p53 (Cawley S et al, Cell 2004). This experiment revealed an unexpectedly large number of TFBS regions, and in particular many of them showed a very particular location, only 22% being located at the 5' end of protein-coding genes while 36% lied within or immediately 3' to well-characterized genes (Fig.15) and were significantly correlated with ncRNAs.

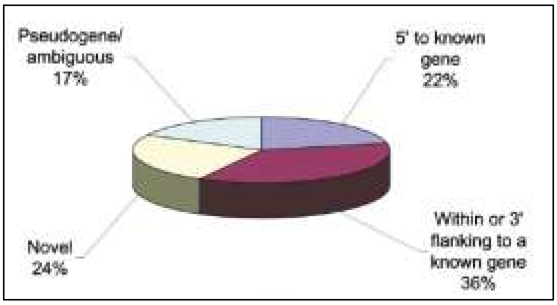


Figure 15. Location of TFBS regions with respect to annotated genes and transcripts on chromosomes 21 and 22 (From: *Cawley S et al, Cell 2004*)**.**

The 5’ located TFBS were generally associated with novel roles for the monitored TFs, while the presence of binding sites outside of known annotations could represent regulatory regions for novel transcripts based on the RNA transcription and EST data, thus making TFBS search a very interesting approach also for the discovery of novel genes, as proved by RT-PCR.

On the other hand, TFBS located within or 3’ to well-characterized genes could either represent distal regulatory elements (i.e. enhancers or silencers) or promoters for non-coding transcripts. In particular, TFBS located into or downstream the last exon were particularly interesting, as they could be involved in interactions with antisense transcripts via the 3’ UTRs of their genes (Lipman DJ, Nucleic Acids Res 1997). Most importantly, all but two of these TFBS were indeed positioned just upstream of possible novel transcripts based either on reported RNA transcription mapping or on mapped ESTs. A clear example was represented by the EWSR1 gene, whose conserved region in the 3’ UTR was consistent with the evidence of antisense regulation. Additional evidence was found by relating these binding sites to full-length mRNAs and ESTs with confidently assignable strandedness, as significant association was found between the proximity to non-canonical TFBS (binding sites not located at the 5’ end of known genes) and the presence of transcription on the opposite strand; more than a half of these sense-antisense pairs were also associated to the conservation of at least one site in mouse genome.

The identified TFBS were also used to extrapolate the putative number of Sp1, cMyc and p53 binding sites on a whole-genome scale; a total of about 12,000 Sp1 sites, 25,000 cMyc sites and 1600 p53 sites were predicted, an amount surprisingly higher than the number of predicted genes itself.

Many pairs of overlapping protein-coding and noncoding RNAs were also found, and quite often these couples were co-regulated (like in the case of the retinoic acid stimulation), suggesting the existence not only of common TFs in their promoter regions, but also the ability to respond to common environmental and developmental conditions. By examining the average correlation obtained during these experiments it was possible to find subpopulations of positively correlated coding and noncoding transcripts, while only a few showed anti-correlation, which would have been expected in the case of gene silencing due to noncoding antisense transcripts. This may suggest the existence of a coordinated expression strategy across the entire genome, or may be achieved via a large-scale chromatin domain that includes both the overlapping transcripts and that is either accessible or inaccessible in a regulated fashion. However, like in the case of protein-coding genes, it seems possible that some (many?) ncRNAs may have biological functions unrelated to those of nearby protein-coding genes.

Finally, although the specific genomic profile of transcription factor binding is probably strictly dependent on the cell types that are being used, there is strong evidence that the behaviour obtained for the three aforementioned TFs will generally apply to other TFs and other cell types.

Redefining the criteria: new approaches to TFBS identification

Due to the significant limitations of existing tools for TFBS prediction, new approaches (often based on a combination of predictive methods) are being developed and two are particularly interesting.

The first approach (Zhu Z et al, Genome Res 2005) is based on the assumption that the functional interactions between TFs often require physical proximity and that, for this reason, their binding sites are likely to be overrepresented in the vicinity of each other. Starting from a TFBS of interest, the algorithm first discovers significantly enriched neighboring motif(s) using human-mouse conserved sequences, then compares these neighbour motifs to known TFBS (obtained from TRANSFAC) in order to understand if they can correspond to new *cis*-regulatory elements, and then moves to examining the functional significance of their physical proximity through the assessment of similarity of their expression profiles (i.e. by analyzing a human cell cycle expression data set). Finally, a confirmation of the accuracy of the prediction can be obtained by comparing the *in silico* results with those obtained in vivo using ChIP-chip technology.

The second approach (Xuan Z et al, Genome Biol 2005) integrates sequence conservation among mammalian genomes and a previously developed promoter prediction program called FirstEF (Davuluri RV at al, Nat Genet 2001) that performs an *ab initio* prediction to identify non-coding first exons and the corresponding promoters. The conservation of mammalian promoter sequences (evaluated to be around 66-68% between human and rodents) associated to FirstEF predictions and transcript information leads to an improvement of 20% and 2% in the specificity and sensitivity of the prediction. Actually the best results are obtained when looking for CpG-island related promoters, by selecting promoter regions defined as -700 to +300 bp around TSSs, but the strongest absolute improvement is associated to non-CpG-related ones. The application of this method is also mainly restricted to known genes, as predicted ones frequently lack complete 5’ end exons.

All the data obtained thus far are stored in the Cold Spring Harbor Laboratory Mammalian Promoter Database (CSHLmpd, http://rulai.cshl.edu/CSHLmpd2).

Personal contribution to the Transcriptome analysis

I started my PhD activity at LNCIB being involved in a functional genomics project producing and analysing human full-length cDNA sequences.

Since the beginning, I found myself very interested in studying the intrinsic information content carried by the human transcriptome. To fulfil this interest I needed first of all to evaluate the many publicly available bioinformatics tools in order to understand their function, selecting and tailoring them to my needs. However, the application of the existing *in silico* instruments highlighted a sum of problems and limitations that quickly pushed me to interacting with informatician colleagues in order to clarify the origin of these restraints, eventually leading to the development and implementation of an integrated computational system (called TSDB), both for an ordered storing of the data and for the scaling-up in the complexity of the executed tasks. This computational system was then used for the high-throughput analysis of cDNA sequences and allowed me, in the end, to define the so-called "LNCIB 5.8K Unique cDNAs collection". In the meantime, I was also able to identify a subgroup of 342 putative novel genes requiring further functional characterization.

My participation to the FANTOM2 (and, lately, to the FANTOM3) project represented a great help to understand how deepening the analysis on these genes. The study of the mouse transcriptome, in fact, clearly draw my attention on the necessity of considering, along with the coding population of the human transcriptome, also events like antisense transcription and, more in general, the involvement of non-coding DNA and of the regulatory mechanisms being involved in these manifestation.

cDNA microarrays experiments (along with other biological assays) were performed with the goal of obtaining a better characterization of the function of these transcripts and gave rise to gene clusters, groups of genes sharing the same level of expression in the examined tissues or cell-lines. The tricky aspect of this analysis consisted in giving a mechanistical characterization to gene clustering results, that is understanding if this behaviour occurred by chance or was regulated by common modulators. For this reason I evaluated the predictive potential of existing algorithms for promoter discovery applied to human genomic sequences, verifying once more the limited applicability of public tools to the required analysis. In particular, these programs seemed to have great problems while performing without *a priori* knowledge. I then exposed my thoughts and my doubts to mathematician colleagues and, in the end, we planned, implemented and validated together a novel algorithm (called ScanPro) that is currently being used for the *de novo* identification of TFBS.

MATERIALS AND METHODS

Extraction of total RNA from tissues

Total RNA was prepared from fresh tissues by modification of the guanidinium–thiocyanate method (Chomczynski P and Sacchi N, Anal Biochem 1987) with adaptation of the cetyltrimethylammonium bromide (CTAB) precipitation method (Del Sal G et al, Biotechniques 1989). Messenger RNA was prepared using an Oligotex mRNA Kit (Qiagen) according to the manufacturer's instructions, starting from CTAB-purified total RNA.

Extraction of total RNA from cell lines

Cell line total RNA was prepared by using the lithium chloride method of Auffray and Rougeon (Auffray C and Rougeon F, Eur J Biochem 1980) with minor modifications. Messenger RNA was prepared using the Oligotex mRNA Kit (Qiagen) according to the manufacturer's instructions, starting from CTAB-purified total RNA.

Full-length-enriched cDNA preparation

First-strand cDNA synthesis reaction was performed using trehalose-stabilized SuperScript II reverse transcriptase (Invitrogen/Life Technologies), which allows for a higher yield of full-length cDNAs (Carninci P et al, Proc Natl Acad Sci USA 1998). The following steps involving the cap-structure biotin derivatization, the capture of the protected mRNA-cDNA hybrids, and the second-strand cDNA synthesis were performed as described in (Carninci P et al, Genomics 1996; Carninci P et al, DNA Res 1997)

Restriction digestion and size selection of double-stranded cDNA

The double-stranded cDNA was digested with the two selected restriction enzymes under standard conditions and subjected to a size selection on an agarose gel. After agarose gel electrophoresis, cDNAs longer than 1500 bp were located using a hand-held long-wavelength UV lamp and quickly excised by using a clean, sterile razor blade. Finally, the Bio-Rad Prep-A-Gene DNA purification kit was used to recover cDNA molecules from the agarose gel according to the manufacturer's instructions.

Ligation of double-stranded cDNA to the vector and propagation of cloned cDNA

cDNA was ligated overnight at 16°C to the chosen vector using 1U of T4 DNA ligase (Invitrogen/Life Technologies). The reaction was performed in a final volume of 10μl in the presence of 50mM Tris–HCl (pH 7.5), 10mM $MgCl_2$, 10mM DTT, 1mM ATP, 25ng/μl BSA.

Reaction was then diluted to 100µl with water, and 2µl of 0.5M EDTA and 2µl of 10% SDS were added. After proteinase K treatment and phenol/chloroform extraction the recombinant plasmid DNAs were purified by using the Microcon YM-100 concentrators (Millipore) bringing the final volume to 10µl in water. Of these, 1µl was electroporated into the host cells, Electromax DH10B (Invitrogen). The number of primary colonies was evaluated by plating different aliquots of the electroporated cells.

Preliminary library quality evaluation

To test the quality of the newly generated cDNA libraries 96 colonies were randomly chosen, picked, and grown overnight in LB medium for subsequent plasmid extraction. Plasmid DNA was obtained using a protocol based on alkaline extraction and binding to diatomaceous earth in the presence of the chaotropic agent guanidinium hydrochloride (Boom R et al, J Clin Microbiol 1990). To check cloning efficiency (by evaluating the percentage of empty vectors) and to determine the average length of cloned cDNAs, these few clones were digested using the cloning restriction enzymes and electrophoresed through a standard 0.8% agarose/TAE gel. To determine the percentage of full-length inserts, 96 clones were subjected to 5V-end sequence analysis.

Bacterial culture growth and plasmid isolation in 96-well format

Clones from selected libraries were grown overnight in 1ml LB broth containing 100µg/ml ampicillin in deep 96-well culture plates. Plasmid isolation was performed in the 96-well format following an optimized protocol based on the alkaline lysis of the bacterial cells and ethanol precipitation of the plasmid DNA.

5'-end sequencing

Starting from plasmid DNA, the insert was cycle-sequenced with fluorescent dye-labeled terminators (Sanger F et al, Proc Natl Acad Sci USA 1977). To obtain 5'-end sequence data a linear amplification of the template was performed using the M13 reverse sequencing primer for pSPORT1 and pCMV-SPORT6 plasmids and the T3 primer for the pBluescript II SK(+) plasmid. The enzymatic dideoxy sequencing reaction was conducted following the manufacturer's protocol with minor modifications. The sequencing products were then precipitated in 60% isopropanol and run on the ABI Prism 3700 automated capillary sequencer (Applied Biosystems).

3'-end and internal sequencing of the 12 unknown potentially coding cDNAs

To obtain 3'-end sequence data a linear amplification of the template was performed using the M13/pUC forward sequencing primer or oligo(dT) anchored primers for those cDNAs in which the poly(A) strand created problems in the linear amplification. To increase the length and significance of the available cDNA sequences, internal primers were designed based on the available 5'-end sequences.

Microarray experiments

1. *cDNA microarray preparation*. Microarray slides were prepared by spotting on Amersham Type 7* slides the purified PCR products of the cDNA clones, diluted in 50% DMSO, using the ChipWriter Pro microarrayer (Virtek). Two different arrays were prepared and used in the present work. The first array accounted for 6336 clones (corresponding to ~5300 unique genes) from the LNCIB full-length cDNA collection and was printed in quadruplicate.

   The first array was used to evaluate gene expression using total RNA extracted from a human tissue (placenta obtained from IRCCS Burlo Garofolo) and from four cell lines, U-118MG (human glioblastoma; obtained from INT, Milan, Italy), OVCAR-3 (human ovarian carcinoma; obtained from INT), CaCo-2 (human colon adenocarcinoma; obtained from INT), and THP-1 (human acute monocytic leukemia cells stimulated with LPS for 12 h; obtained from EMBL, Heidelberg, Germany). The second array accounted for 8160 cDNA clones from the IMAGE collection (Research Genetics/Invitrogen) and for the 6336 LNCIB full-length cDNA clones and was printed in duplicate. This second array was used to obtain gene expression profiles from 30 human advanced ovarian cancer samples and from the four cell lines, obtained by various treatments of the original IOSE-hTERT_INT cell line (an SV40-immortalized normal ovarian surface epithelium line stably transfected with the human telomerase cDNA (hTERT), kindly provided by Dr. S. Canevari, Department of Experimental Oncology, Istituto Nazionale Tumori of Milan, Italy). In particular, in addition to the original IOSE-hTERT_INT cell line, the following three sublines were investigated: the first obtained after 5 weeks 5μM LiCl treatment of the original cell line (LiCl), the second obtained after 5 weeks treatment of the original cell line with 10ng/ml PDGF and 10ng/ml FGF2 (FGF2/PDGF), and the third obtained after 5 weeks treatment of the original cell line with 10ng/ml FGF2 and maintenance in RPMI (RPMI_FGF2).

2. *cDNA target preparation*: *study specimens RNA and reference RNA*. For all experiments described total RNA was extracted from cultured cells or tissue using TRIzol reagent (Invitrogen) according to the manufacturer's recommendations.

   Microarray targets were prepared by following the two-step indirect fluorescent labeling method (Xiang CC et al, Nat Biotechnol 2002) starting from 10μg total RNA. First-strand cDNA synthesis reaction was performed using random primers in the presence of aminoallyl–dUTP and the resulting amino groups were then coupled to the NHS–ester of the fluorescent dye. In the case of the first array all RNA samples were separately labeled with both Cy3 and Cy5 dyes and used in self-to-self hybridizations, while in the case of the second array tumor samples and the four IOSE-hTERT_INT sublines were labeled with Cy5, while the Cy3 dye was used to label the reference RNA, obtained by mixing equal amounts of human placenta, U-118MG, OVCAR-3, CaCo-2, and THP-1 RNA.

3. *Hybridisation and slide washing*. Hybridizations were performed in the presence of 3xSSC, 0.2% SDS under coverslips as specified by the aforementioned fluorescent labeling method, by overnight incubation at 63°C using a manual hybridization chamber.
4. *Slides scanning and image analysis*. After hybridization and washing, slides were scanned using the GenePix 4000B microarray scanner (Axon). Images were acquired for Cy3 and Cy5 channels in a 16-bit TIFF format and then analyzed using GenePix software (Axon). All images were manually flagged to remove artifacts and bad spots. Only unflagged features displaying a mean fluorescence intensity greater than the local background plus 3 standard deviations and with that level of expression confirmed in both channels and in all the replicated measurements were considered as expressed in each self-to-self hybridization performed with human placenta, U-118MG, OVCAR-3, CaCo-2, and THP-1 RNA samples. In the case of expression profiling experiments only unflagged features displaying background-corrected positive intensities were considered. All the experiments were normalized using the Lowess method and a maximum of 10% missing values was allowed in the analysis.

Cell culture and transfections

U2OS cells (osteosarcoma cell line, with p53wt) were routinely cultured in Dulbecco's modified Eagle's medium (Invitrogen) supplemented with 10% fetal bovine serum, penicillin (100U/ml), and streptomycin (100µg/ml). For the experiments, cells were grown on coverslips in 35-mm petri dishes containing 8 x 10^4 cells per dish. After a 24-h incubation at 37°C in 5% CO_2 atmosphere, cells were transfected using the calcium phosphate precipitation method.

Transfections were performed using 8 of the 12 cDNAs, cloned into the pCMVSport6 (Invitrogen) expression vector. To detect the transfected cells the green fluorescent protein was used in a 1:10 ratio with respect to the DNA of interest.

For clone 5000FJE08 anti-cytochrome c mAb (Promega) was used to study the mitochondrial localization and biotinylated concanavalin A (Roche) to investigate the endoplasmic reticulum.

For clones 5000AFE01 and 5000EDD03 anti-human nucleolin mAb D3 (Deng JS et al, Mol Biol Rep 1996) was used to study the nucleolar localization.

Actin filaments were detected in clones 5000CIC06, 5000GCA01, and 5000GDH05 using TRITC phalloidin (Sigma Chemical Co.).

Twenty-four and 48 h after transfection the cells were examined by epifluorescence with a DMLB Leica microscope or a Zeiss laser scan microscope (LSM510) equipped with a 488λ argon laser and a 543 λ helium neon laser.

Northern blot analyses

Northern blot analyses for the unknown transcripts of interest were performed essentially as described by (Hla T and Maciag T, J Biol Chem 1990). cDNA inserts (~500 nt) corresponding to the 5' end of the unknown clones of interest or GAPDH were labeled to a high specific activity ($<10^8$cpm/µg DNA) by using a random primer labeling kit (Invitrogen). Membranes were then hybridized, extensively washed as described in the aforementioned manuscript, and subjected to autoradiography. Typically, the membranes were exposed at -80°C overnight.

RESULTS

The first part of the work presented in this thesis is related to the LNCIB full-length cDNA project, more specifically on the high-throughput analysis of human cDNA libraries and the identification of the non-redundant collection of unique transcripts to characterize the subset of novel genes and their subsequent functional analysis. Many publicly available bioinformatics tools were evaluated and in the end an integrated computational system (called TSDB) was implemented to efficiently manage the analysis and annotation processes.

PART 1.1 INTEGRATED BIOINFORMATICS PLATFORM FOR THE ANALYSIS OF THE TRANSCRIPTOME

With the amount of genomic information now available, new software systems need to be integrated to approach the many analytical steps into high throughput pipelines. A particularly interesting field for this kind of application is the analysis of nucleotide sequences obtained from automated DNA sequencers from the analysis of cDNA libraries (Fig.16).

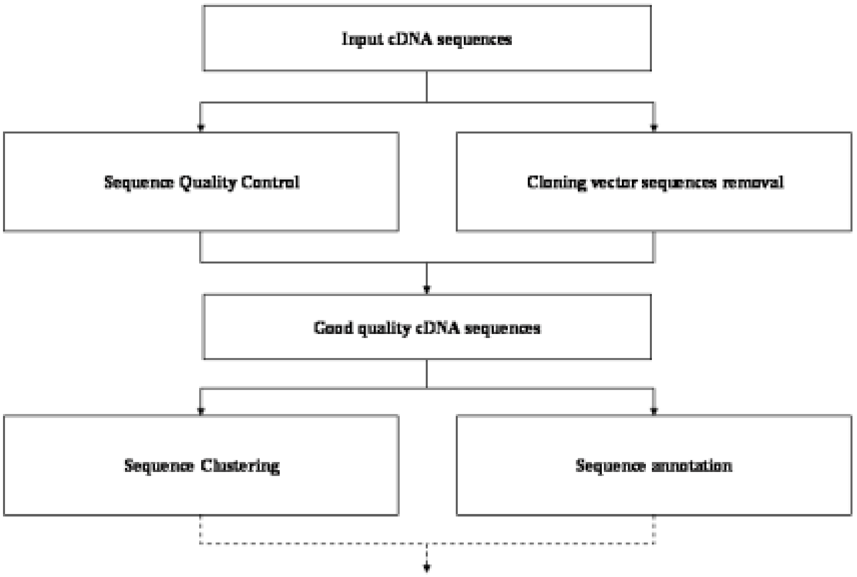


Figure 16. General scheme of cDNA sequence analysis workflow.

The chromatograms generated by the sequencer need to be checked in order to verify their quality, as bad quality and cloning vectors sequences need to be removed. This can be done by simply recording the base calling at a nucleotide level, or can be refined by introducing more accurate evaluation criteria defined by the operator.

Along this, the complexity of a cDNA library is generally quite limited, thus leading to the necessity of identifying all the unique cDNAs available, including (if necessary) the putative isoforms of each gene. This goal can be achieved in two ways: by a direct comparison across all the sequences, or in an indirect way, by assigning each cDNA a functional annotation, using this information as a filtering rule. In the first case the user must select a clustering algorithm and perform the assembling, in the second case a local alignment tool is required in order to define the regions of homology between the examined sequences and reference sequences which already possess a functional annotation. These two methods can also be used in parallel in order to strengthen the final result.

Many public bioinformatics tools are available for the aforementioned specific activities and they were all tested in order to choose the most suitable for the analysis pipeline.

Phred (Ewing B et al, Genome Res 1998) is the most used open source instrument for evaluating sequence quality and its choice was obvious, as well as the choice of Cross_match (P. Green, PHRAP and CROSS_MATCH, University of Washington, Seattle, WA, USA; unpublished) for cloning vector removal. No other open source tool, indeed, could have ever been chosen to perform these tasks. The same thing can also be said for sequence annotation, as BLAST is the most used local sequence alignment algorithm and many reference databases can be queried to study different aspects of cDNA sequences. The most tricky choice concerned sequence clustering, as many algorithms existed: StackPACK (Burke J et al, Genome Res 1999) was chosen as it was the only one that allowed incremental clustering, that is the possibility to compare new sequences with already analyzed ones without the need of disrupting existing gene clusters. Other specialized tools required during the different phases of cDNA sequence analysis, such as the one needed for exon-intron boundaries definition or for the identification of coding regions located into unknown transcripts, were evaluated and selected after an appropriate comparison to other (if any) similar.

Despite the existence of all these resources, integrating (in an open-source package) heterogeneous analysis tools is a difficult task. As a matter of fact, only a few of such tools exist (Shah M et al, http://compbio.ornl.gov/tools/pipeline 2000; Taudien S et al, Trends Genet 2000) and most of them do not contain a database component where information can be stored and easily accessed. Moreover, complex queries are limited within traditional database applications (Gupta A, SIGMOD RECORD 2004): the level of interaction that the few existing databases allow is fairly limited, enabling only a narrow range of analyses (Inman JT et al, IBM System Journal 2001). Finally, and most importantly, although queries could benefit from the availability of temporal information, no system is adequately equipped for time management. At best, available tools provide data with a version number and the date of the last update, but they do not allow to retrieve complete information with respect to their evolution over time. Along this line, available databases often undergo radical changes, due to the great amount of new data (that can also possibly introduce assembly and/or human manual errors), and the availability of new releases becomes the source of relevant changes in information associated with a sequence. Hence, in order to fully understand the characteristics of clones/sequences from their annotation, it is necessary to record the history of each clone to trace the variations it undergoes.

Integrated computational system for the temporal analysis of nucleotide sequences

Given the above described problems, I pushed the decision to develop an integrated computational system (Braidotti M et al, reviewed manuscript re-submitted to BMC Bioinformatics; see also *Personal Publications*) that supports the production and temporal analysis of biological sequences, focusing on nucleotide sequences obtained from the production and sequencing of full-length human cDNA libraries.

This system consists of three main components: the software pipeline, that filters and analyzes the generated sequences, the temporal sequence database, that stores the data as well as the results of data analysis, and the user interface, that allows to query the database and to export its contents.

1) *The software pipeline*. It consists of a series of Perl scripts which interact with the file system, the temporal database, and the internal and external analysis tools. Its behaviour can be summarized as follows. First, it receives a set of chromatograms generated by the sequencer as input, and it performs a quality control for the removal of bad quality sequences and of other unsatisfactory sequences (anomalous bases repetitions, too short sequences, etc.), keeping track of the outcomes of such an analysis. Successively, the selected sequences undergo the so-called "vector trimming" phase, which identifies and removes the bases belonging to the cloning vector. The next phase is the annotation phase which consists of two main steps: the clustering step and the functional annotation step. The clustering step compares each sequence to all the others, looking for similarities that allow one to identify sets of sequences which correspond to the same gene (with the possibility of detecting alternative splicing forms). The functional annotation step compares all the generated sequences with those belonging to various public databases, with different degrees of functional significance, looking for local similarities between the query sequence(s) and the database ones. These two steps are somehow independent, and thus they can take place in either order. Moreover, they can also be executed on groups of FASTA-formatted sequences, with no chromatrograms available, obtained from external sources.

   If no available external tools exist for a given analysis task, the pipeline uses the internally developed software modules to perform such a task. All relevant information are stored into the temporal database. When only incomplete data are available one can force the process to start at a different entry point by setting a parameter of the pipeline activation script. This is very useful when someone is analyzing sequences that have not been produced internally, such as those obtained from external laboratories or directly downloaded from a public sequence database.

Phred. First of all sequence quality is estimated by Phred (Ewing B et al, Genome Res 1998), which provides the probability of error for each base. A quality file is generated for each sequence read and every file is checked to keep only sequences with a sufficient number of good-quality bases. Q20 or higher accuracy was considered successful, where Q20 means one possible error in 100 bases.

Cross_match. Then Cross_match (P. Green, PHRAP and CROSS_MATCH, University of Washington, Seattle, WA, USA; unpublished) is run on the good sequences left. This software masks the cloning vector sequence situated before the beginning of the cDNA sequence. Masked bases are then removed automatically to keep only insert sequences.

Quality control. To improve the overall quality of the sequences a second quality check is made. cDNA sequences whose length is less than 80nt are removed by an internal Perl module, as well as sequences with more than 25 contiguous repetitions of the same nucleotide (Poly-C and Poly-A excepted) or sequences with an excessive abundance of a single nucleotide (more than 48% of the total length of the sequence).

StackPACK. After filtering, an incremental clustering procedure is applied to the data to identify all the unique genes and their possible alternative splicing forms and, where possible, to select full-length representatives for each gene. This is done using StackPACK, a software based on the d2_cluster algorithm (Burke J et al, Genome Res 1999). By using PHRAP (P. Green, PHRAP and CROSS_MATCH, University of Washington, unpublished) the sequences grouped together by d2_cluster are aligned and assembled and alignment quality is improved by removing particularly distinct sequences: clusters are generated if at least two sequences share a common region of a given length (100 bp in this case). All sequences not belonging to any cluster are considered as “singletons”. For each cluster the best representative sequence (i.e., the 5’ more complete, longest sequence starting with the poly(C) stretch introduced during the cDNA synthesis) is chosen along with one or more alternative splicing forms, when present.

BLAST queries. Sequence similarity searches are executed by using NCBI BLAST (Altschul SF et al, J Mol Biol 1990). The BLAST search looks for matches in RefSeq, Unigene, or nr databases. For each query, up to 5 alignments are stored, depending on the e-values associated to the results (fixed threshold: 1e-40 for RefSeq; 1e-50 for Unigene; 1e-60 for nt). These thresholds have been chosen empirically, depending on the stringency I decided to apply to my data, but they can be easily modified to user discretion. The number of stored alignments has been fixed assuming 700nt as the average length of cDNA sequences, and 150nt as the average length of human exons. In this way, it should be possible to store information about all the exons of the cDNAs, if they should give different High Scoring Pairs (HSP) in BLAST output. It is clear, however, that similarity searches usually do not provide "certain" results, so significant matches never imply a "certain" functional annotation. This is another reason for storing information in a temporal way and for storing 5 different results that can be compared one another to verify cDNAs annotation. Matches from RefSeq are considered the best for the functional enrichment of cDNAs annotation, those from nr the worst. If there is no BLAST result fulfilling the e-value threshold, the sequence is marked as unknown.

BLAST analysis can also be made using other databases. In particular, searches on GenBank Human EST and Human Genomic databases are performed to retrieve additional information on, for instance, the genomic location of the sequences and the annotated elements, e.g., predicted genes and transcripts, proteins and orthologous matches, located in the same genomic region.

2) *The temporal sequence database*. The Temporal Sequence DataBase (TSDB) is a fundamental component of the system. It records data related to sample plans, sequences and sequence analyses. In particular, it stores the results of the operations of sequence annotation (via BLAST) and rearraying. In fact, it allows to record multiple BLAST and clustering results, the first ones obtained by querying different public and private databases as well as by querying the same database at different times, the second ones obtained by multiple executions of possibly different clustering algorithms. The distinctive features of TSDB are its flexibility, that allows to deal with incomplete information, and the presence of a temporal dimension, which makes it possible to keep track of data history. Another relevant feature of TSDB is its user-friendly nature. It provides a number of predefined queries that allow the user to properly integrate information distributed over different tables (i.e. the user can retrieve all the clones belonging to a certain library, with the same clusterID, and whose alignment score difference is 5% maximum). Moreover, specific additional queries can be easily formulated and executed. The last functionality of TSDB to be mentioned is its support to information sharing and exchange. Additional data, such as new information about genomic location, ontologies, or orthologous comparisons, possibly generated by other experiments carried out in the laboratory (e.g., cDNA microarrays gene expression experiments), can be easily integrated into TSDB. In a similar way, relevant data contained in TSDB can be easily exported to other databases.

TSDB is based on a temporally-extended data model obtained by adding suitable temporal features to the Enhanced Entity-Relationship (EER) model (Elmasri R and Navathe SB, Fundamentals of Database Systems 2004). This temporal data model integrates the basic constructs of the EER model with temporal primitives borrowed from the RAKE model (Ferg S, Proceedings of the 4th International Conference on the Entity-Relationship Approach 1985) and the TimeER one (Gregersen H and Jensen CS, TimeCenter Technical Report TR-35 1998). The TSDB temporal data model describes the evolution of data as a sequence of database states *A,B,C, . . .*. The transition from any given state *A* to the next state *B* is determined by the occurrence of a specific update *X*. Hence, for any sequence of states *A,B,C, . . .* there exists a sequence of updates *X, Y,Z, . . .* such that $A \xrightarrow{X} B \xrightarrow{Y} C \xrightarrow{Z} \ldots$. States persist over time, that is, they hold over a time period, while updates take place instantaneously, that is, they occur at time points. The time period during which a given state is current in the database is identified by a pair of timestamps: BEGINstamp and ENDstamp. BEGINstamp and ENDstamp are the occurrence time of the update that initiates the validity of the state and the occurrence time of the update that terminates it, respectively. Each timestamp is represented by a pair (*date, time*) that associates a time value with a data value (they correspond to the date and time elementary temporal domains of SQL).

The TSDB temporal data model encompasses all the basic constructs of the EER model, namely, entity and relation types, attributes, internal and external identifiers, participation and cardinality constraints associated with entity-relation links (using the (*min, max*) notation), and generalization/specialization hierarchies. States are modelled by temporal relation types which are obtained by adding the BEGINstamp and ENDstamp timestamps to basic relation types. BEGINstamp and ENDstamp timestamps can also be used to model time-varying attributes, that is, attributes whose value can change over time. Updates are not explicitly represented. One of the most sophisticated temporal features of the TSDB model is the possibility of distinguishing between snapshot and lifespan constraints associated with entity-relation links. Given an entity *E* and a relation *R*, it is possible to label the *E-R* link with a *snapshot constraint* (*minS, maxS*) to state that, at any time point, each instance of *E* participates in at least *minS* and at most *maxS* instances of *R*. Moreover, it is possible to label the *E-R* link with a *lifespan constraint* [*minL, maxL*] to state that, over time, each instance of *E* participates in at least *minL* and at most *maxL* instances of *R*. It obviously holds that $minS \leq minL$ and $maxS \leq maxL$.

This conceptual schema was translated into a relation one, which was implemented in MySQL. MySQL is a simple, fast, and stable DataBase Management System (DBMS), which scales well. It supports most of the distinctive database features and it can be easily paired with the scripting languages used to implement the software pipeline and the interface.

3) *The user interface*. The interface supports three different kind of search. The first one (Reporter) allows to obtain data on the status of the sequences. In particular, it allows the sorting of sequence IDs associated with sequence text, PHRED quality file, direction of sequencing, mapping location (container and fridge info), analysis status, Poly-C position and GenBank link (whenever the sequence has been submitted to NCBI). The second search (BLAST Results) allows the display of results from queries delivered to the GenBank RefSeq, Unigene, nr, Human EST, and Human Genomic databases. For each sequence, it is possible to obtain the corresponding NCBI accession number (with the link to the specific page of NCBI's Entrez), the annotation, and its degree of full-lengthness with respect to the database entries. In addition, data can be queried, or filtered, by exploiting the GUI tools such as filter by selection and filter by form. The last search (Clustering) allows to display two different types of information about a given sequence collection: the list of unique "genes" (singletons or cluster consensus sequence) that are present in it and, for each sequence, the cluster, or the singleton, it belongs to. For every query, a mask can be applied to restrict the sorting to a subset of the sequences that satisfy user's criteria.

[Pipeline software and temporal database can be downloaded from httb://www.bioinfo.lncib.it/]

PART 1.2 EXPERIMENTAL TESTING

Library Selection

All the libraries were produced with the CAP trapper method to maximize the number of 5' end full-length cDNAs.

After a preliminary phase of library quality evaluation (see Materials and Methods), four libraries constructed from RNA extracted from the MOLT-4 cell line (human T cell acute lymphoblastic leukemia), from human placenta tissue (39 weeks fresh placenta tissue from cesarean delivery), and from two distinct human fetal brain tissues (respectively 15 and 18 weeks from right frontal lobes) were selected for further sequencing characterization.

During the sequencing phase ~23,000 5'-end sequences were generated: ~9000 from the MOLT-4 library, another ~9000 from the 15-weeks fetal brain library, ~3500 from the placenta library, and ~1500 from the 18-weeks fetal brain library.

Integrated computational system analysis: data filtering

By introducing the 23,000 initial sequences into the developed integrated computational system (Fig.17) it was possible to remove those that, for various reasons, did not match quality requirements.

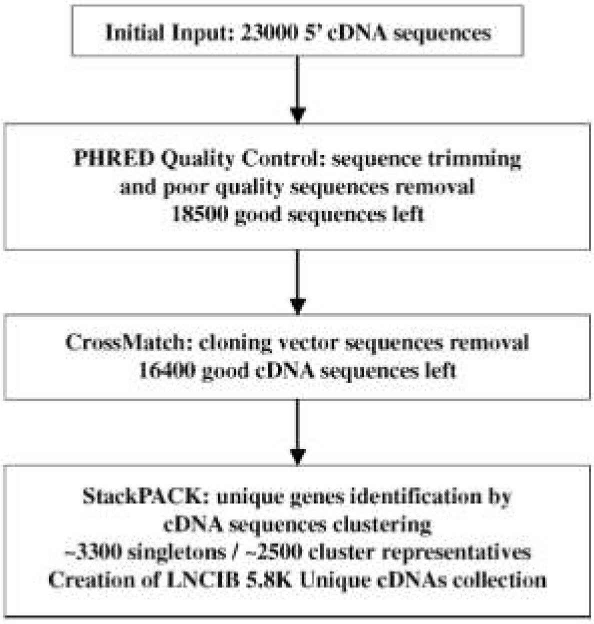

Initial Input: 23000 5' cDNA sequences
PHRED Quality Control: sequence trimming and poor quality sequences removal 18500 good sequences left
CrossMatch: cloning vector sequences removal 16400 good cDNA sequences left
StackPACK: unique genes identification by cDNA sequences clustering ~3300 singletons / ~2500 cluster representatives Creation of LNCIB 5.8K Unique cDNAs collection

Figure 17. Integrated computational system analysis. Various analytical tools (both public and internally developed) were used to analyze sequence quality (Phred and Cross_match) and to find unique genes among the cDNA collection (StackPACK).

After bad-quality sequences, empty cloning vectors and artifacts were removed, 16,400 good-quality cDNA sequences were recovered and this group was used as input for the clustering procedure. StackPACK generated ~2100 clusters with ~3300 sequences remaining as singletons. By choosing the best representative sequence for every cluster I found that, in some cases, two or more sequences had to be chosen, due to alternative splicing events. For this reason the ~2100 clusters generated ~2500 cluster representatives. This set of ~5800 sequences was then considered for further analysis to obtain a functional annotation for every sequence (Dalla E et al, Genomics 2005; see also *Personal Publications*).

Integrated computational system analysis: data recording

All the information belonging to the cDNA molecules clustered into the "LNCIB 5.8K Unique cDNAs collection" were stored into TSDB. Along with nucleotide sequences and analysis results, also information on technical aspects of libraries production such as sample plans, containers, and enzyme tables were recorded as relevant.

As previously explained in Materials and Methods, one of the main features of TSDB is the fact of being a temporal database. This means that it allows to record the results of the execution of the same query at different time points, making it possible to detect the temporal evolution of the available knowledge about the features that characterize every sequence. Although results based on obsolete information should simply be discarded, since they are no longer supported, I do not consider this to be the best solution.

Databases, in fact, undergo changes that need to be verified over time, as they are not always correct. The most common problems are small mistakes, like swapping the annotation of different isoforms of a given gene. For example, the initial annotation of a clone was "*Homo sapiens* IMP (inosine monophosphate) dehydrogenase 1 (IMPDH1), transcript variant 1" and was lately replaced by "*Homo sapiens* IMP (inosine monophosphate) dehydrogenase 1 (IMPDH1), transcript variant 2". While at a first glance this could represent a little change, it is indeed quite relevant since different members of the same gene family, as well as alternative splicing forms of the same gene, can perform very different functions and be involved in very different biological behaviours. However more striking errors occur, too, and lead to the removal of the entry from public databases, like in the case of another cDNA whose first annotation was "*Homo sapiens* apoptosis inhibitor (FKSG2)", located on chromosome 8p11.2, and that lately resulted to be "*Homo sapiens* FLJ44635 protein", located on chromosome Xq13.1.

Overall, 1281 clones out of 20,266 submitted to at least two rounds of analysis underwent changes in their annotation since their recording into TSDB. In particular, 146 of these clones changed their LocusLink IDs. Another relevant group of 500 clones could not be compared because their GenBank accession number changed, the LocusLink ID being available either for the first or for the second round of analysis, but not in both cases. Finally, 1600 clones matched to different GenBank accession numbers, but shared the same common LocusLink ID, suggesting the same gene.

It is quite clear that, when dealing with cDNA microarrays gene expression experiments and TFBS search, it is not possible to accept this kind of mistake and it is also not possible to wait until an update of the database is released. For this reason, the temporal evolution of the sequence features becomes the best way of solving this kind of constraint, and this is the reason why, whenever new cDNA sequences are generated, a new and complete BLAST analysis is performed to update the cDNAs annotations. Similarly, a new analysis is also performed in the case of the release of a new version of one of the GenBank databases.

Identification of known genes

The 5295 clones that were identified as singletons or unique cluster representatives at the end of the clustering step were submitted to a BLAT (Kent WJ, Genome Res 2002) analysis for the identification of clones corresponding to known sequences.

Altogether, ~79% of the unique sequences displayed significant matches with entries from RefSeq db (human division) with another ~11% of the sequences recovering significant matches when GenBank, EMBL, and DDBJ human mRNA entries were also considered. The last ~10% remained "unknown".

In those cases where I observed that the similarity between the clone and the mRNA sequence did not involve the whole clone sequence, I mapped both sequences on the human genome (assembly 34). Only pairs mapping at the same genomic location were considered. There were 4137 clones that were identical or had a different 5' length compared to the known sequence, while 662 sequences were likely produced by alternative splicing of a known gene.

Comparing the 5' ends of LNCIB clones with the reported 5' ends of RefSeq mRNAs I observed that 1769 of the analyzed clones (more than 40%) extended the previously annotated 5' end, 1238 (~30%) proved to be as long as the known sequences, and 292 (~7%) were a maximum of 200 nt from the reported 5' end (the majority conserving the start codon), while a significant fraction, 884 (~20%) of them, were more than 200 nt shorter (Fig.18).

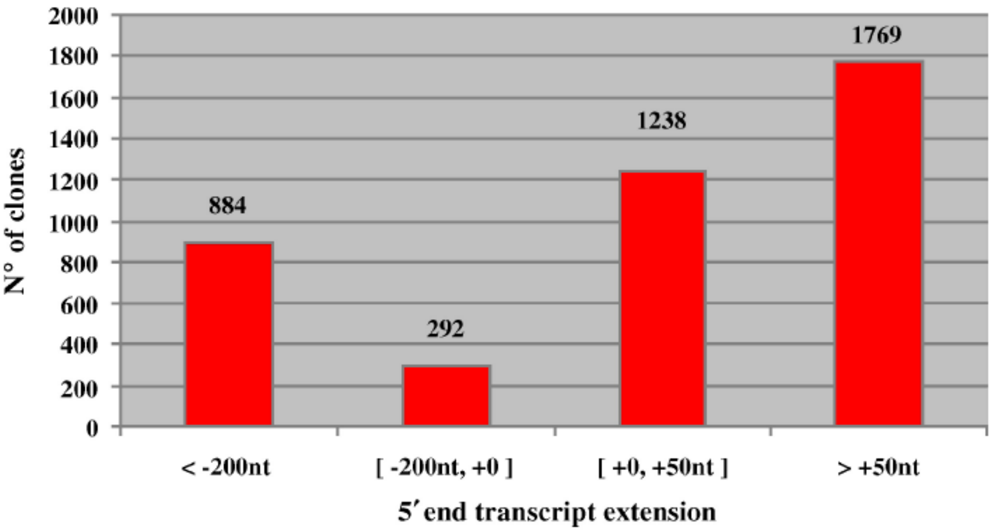


Figure 18. Distribution of length differences of 5' end between LNCIB and RefSeq transcript sequences.

Characterization of unmatched clones

The remaining subgroup of 496 clones (Fig.19), corresponding to putatively unknown clones, was submitted to a BLAT query on the human genome sequence, and for 417/496 clones I identified their genomic location.

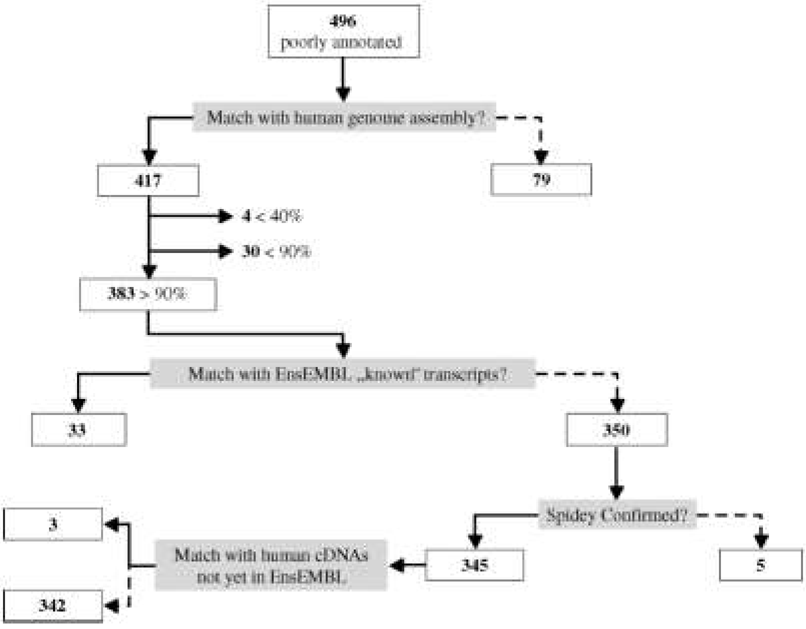

496
poorly annotated
Match with human genome assembly?
417
79
4 < 40%
30 < 90%
383 > 90%
Match with EnsEMBL „known" transcripts?
33
350
Spidey Confirmed?
345
5
Match with human cDNAs not yet in EnsEMBL
3
342

Figure 19. Human genome BLAT analysis. Flowchart showing the analysis pipeline followed for the genomic characterization of the 496 poorly annotated clones. Dashed arrows mean negative answer. The Spidey program was applied comparing the clone sequence and the genomic sequence of the matching EnsEMBL transcript.

Only the 383 clones with a genome match coverage of >90% were included in the second stage of the analysis pipeline.

By comparing the genomic coordinates of these clones with the features annotated in EnsEMBL (Rel 20 based on NCBI Rel 34, http://www.ensembl.org/Homo_sapiens/) I identified 33 more clones matching with EnsEMBL “known” transcripts. These transcripts are termed known since they are supported by the presence of full-length cDNAs or protein sequences already annotated in public databases and have thus been assigned an established gene symbol.

To investigate further the remaining 350 clones I compared their sequence with the corresponding genomic sequence using the Spidey program (Wheelan SJ et al, Genome Res 2001) (http:// www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/). Spidey has been specifically designed to compare mRNAs to genomic sequences, mapping the genomic coordinates of exons, and checking for the presence of canonical donor/acceptor splice sites. I excluded from further investigation 5/350 clones that showed a genome match at multiple locations without canonical splice sites. Of the remaining 345 clones fully matching with the genomic sequence, 307 showed a single match and 38 multiple matches separated by canonical splice sites.

This set of 345 clones, possibly representing novel genes, was then subjected to further bioinformatics and expression analyses. Among the features annotated in EnsEMBL are matches with human cDNAs, UniGene and dbEST entries, proteins, Genscan predictions, and other vertebrate mRNAs. Three of 345 clones matched with human cDNAs not yet present in the RefSeq database, reducing the number of potentially novel genes to 342: 214 obtained from the human fetal brain libraries, 114 from the MOLT-4 cell line, and 14 obtained from the human placenta library.

Functional annotation by comparison with human ESTs and other vertebrate transcripts

Of the 19/342 clones showing a significant match with other vertebrate mRNAs, 16/19 also matched a predicted Genscan gene and 13/19 hit human ESTs, strongly suggesting that these may represent genuine novel human genes. Among the other 323 clones (33 with a spliced match), 155 showed a match against human ESTs and 81 of these also matched with a predicted Genscan gene. Of the remaining 168 clones (12 with a spliced match), 76 showed a match with a Genscan prediction. Finally, 92 clones did not show any match with EnsEMBL features, with only 5 of them showing a spliced match with the genome.

Expression analysis

cDNA microarray self-to-self hybridization were performed by using four different cell lines, U-118MG (glioblastoma cell line), OVCAR-3 (human ovarian carcinoma), CaCo-2 (human colon adenocarcinoma), and THP-1 (human acute monocytic leukemia cell line stimulated with LPS), and a human tissue, placenta, as sources of mRNA molecules for studying the expression levels (if any) of the spotted cDNAs. All RNA samples were separately labeled with both Cy3 and Cy5 dyes and used in self-to-self hybridizations on a cDNA microarray slide, accounting for the 6336 different physical clones spotted in quadruplicate. In particular I focused on the group of 342 EnsEMBL unknown clones to evaluate whether these clones could represent transcribed sequences. I selected only those genes with a signal greater than the background intensity plus 3 standard deviations and with that level of expression confirmed in both channels and in all the replicated measurements.

Fig.20A shows the scatter plot of the 260 unknown transcripts expressed in the investigated samples along with the expression values measured in placenta for 6 housekeeping genes. Fig.20B shows a portion of the hybridization performed on the LNCIB 5.3K unique cDNA collection by using cDNA target synthesized from placental RNA.

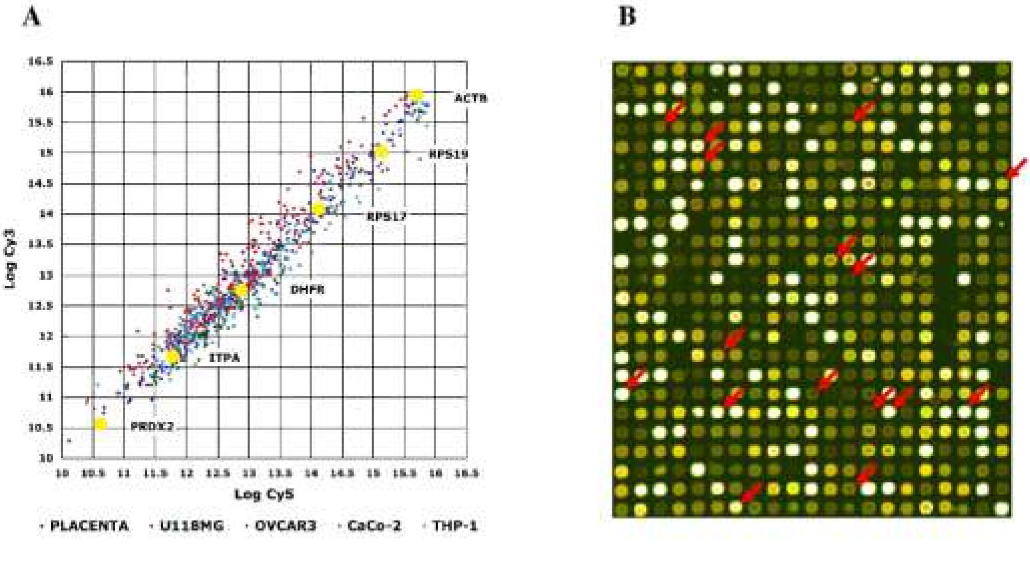
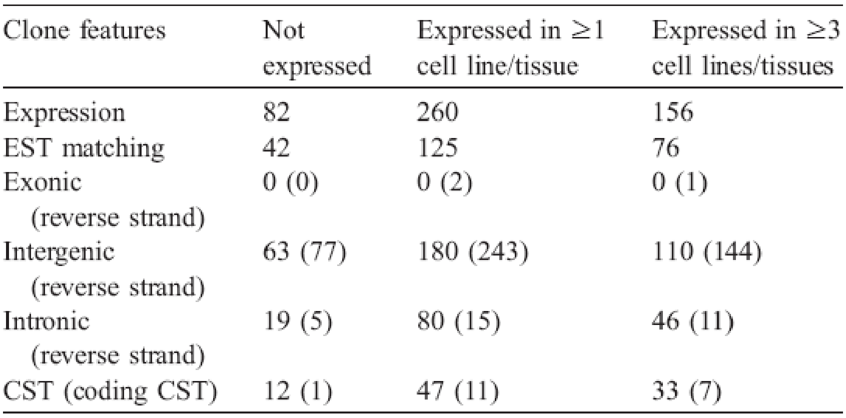


Figure 20. DNA microarray gene expression analysis. (A) Scatter plot of the unknown transcripts expressed in the following investigated samples: human placenta, U-118MG cell line, OVCAR-3 cell line, CaCo-2, and THP-1 cell line. Cy3 and Cy5 values are expressed in logarithmic scale (base 2). Expression values measured in placenta for 6 housekeeping genes are shown as yellow circles (PRDX2, peroxiredoxin 2; ITPA, inosine triphosphatase (nucleosine triphosphate pyrophosphatase); DHFR, dihydrofolate reductase; RPS17, ribosomal protein S17; RPS19, ribosomal protein S19; and ACTB, β-actin). (B) Snapshot of the hybridization performed on the LNCIB 5.3K unique cDNA collection by using cDNA target synthesized from placental RNA. Red arrows point to 16 expressed unknown transcripts.

Table 1 summarizes the results on the subgroup of 260 clones expressed, at different levels, in all five experiments: 170 clones belong to the human fetal brain library, 79 from the MOLT-4 library, and 11 from the human placenta library.



Clone features	Not expressed	Expressed in ≥1 cell line/tissue	Expressed in ≥3 cell lines/tissues
Expression	82	260	156
EST matching	42	125	76
Exonic (reverse strand)	0 (0)	0 (2)	0 (1)
Intergenic (reverse strand)	63 (77)	180 (243)	110 (144)
Intronic (reverse strand)	19 (5)	80 (15)	46 (11)
CST (coding CST)	12 (1)	47 (11)	33 (7)

Table 1. Expression features, genome location according to the EnsEMBL annotation in the forward (and reverse) strand and **occurrence of conserved sequence tags (CST)** in the comparison with the mouse genome of the 342 clones likely corresponding to novel human genes

About 60% of these clones (156/260) are expressed in more than three cell lines/tissues. These clones were mapped with respect to EnsEMBL genome features annotated both on the forward and on the reverse strand and the majority of them are located in intergenic regions; in particular, 107 are located at least 10 kb from the nearest EnsEMBL transcript, thus possibly representing completely novel genes, while 17 more cDNA clones have a distance of at least 5 kb from the nearest EnsEMBL transcript. The exact meaning of these results, in particular for the cDNAs that are nearer to known genes or transcripts, could be explained considering that new transcriptional units can be located at the very end of known genes (Cawley S et al, Cell 2004) or that the CAGE technique (Shiraki T et al, Proc Natl Acad Sci USA 2003) has shown that new TSP can be found also 10 kb or more from the established 5' end of a known gene. The remaining clones, mapping in intron regions of already annotated genes, could alternatively be representatives of novel gene splicing isoforms. Interestingly, 2 clones showed antisense matches with annotated exons.

By looking in more detail it was finally possible to identify subgroups of genes that appear to be expressed or not expressed in a specific way and in particular different subgroups presenting a significant degree of co-expression in the different targets analyzed.

In addition gene expression profiling was performed on 30 advanced epithelial ovarian cancer specimens and on several cell lines, after various treatments of the IOSE-hTERT_INT cell line (see Materials and Methods for details). This analysis showed that 289 of 342 unknown clones survived the filtering procedures, being differentially expressed in the samples considered (Fig.21A). This is very important as it shows how in a microarray experiment, by using RNA sources that are different from those used for library construction, it is still possible to identify new expressed genes previously unidentified. To group the expressed genes in meaningful clusters it will be necessary to increase the number of available ovarian cancer specimens. By now, it is not yet possible to understand the function of the unknown clones by taking into consideration the known genes that cluster together. As outlined in the second part of my thesis, this problem pushed my interest in analyzing and developing tools for TFBS identification to identify conserved elements putatively responsible of gene co-regulation and co-clustering.

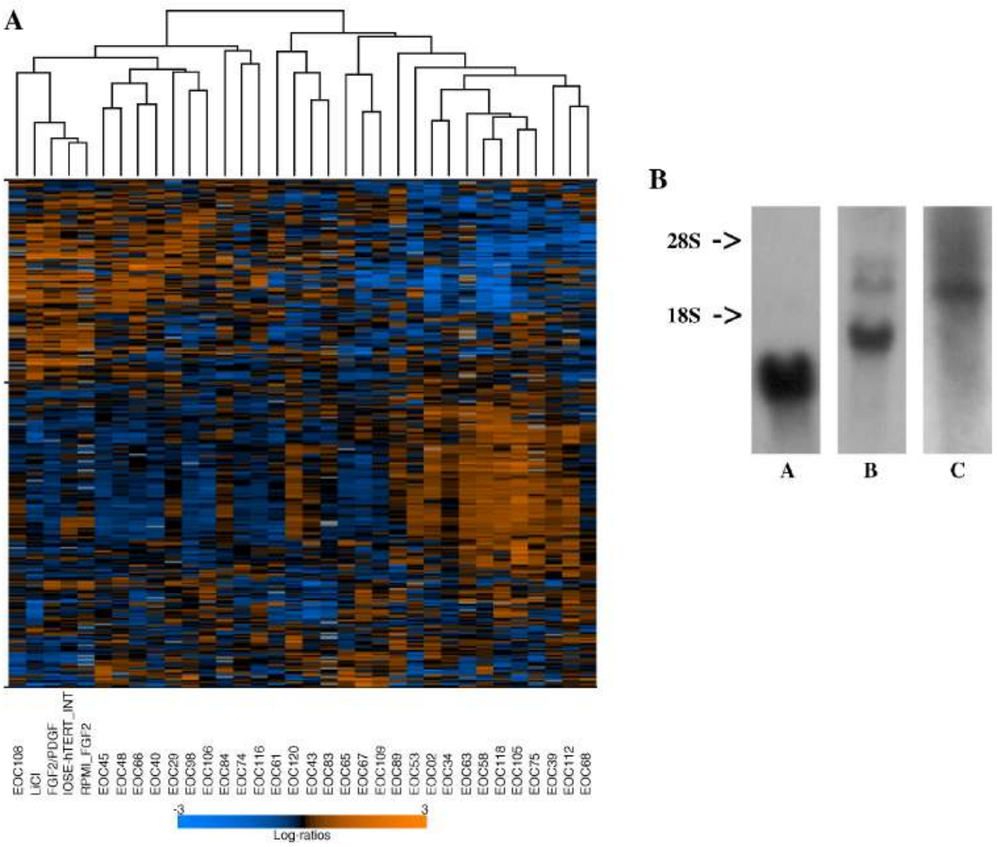


Figure 21. (A) Gene Clustering. Gene expression pattern determined using agglomerative hierarchical clustering of 30 primary ovarian cancer samples and four cell lines using the 289 unknown clones that survived the filtering procedures. The Pearson's metric and the average clustering method were used. Expression levels are relative to a common reference obtained by pooling equal amounts of human placenta, U-118MG, OVCAR-3, CaCo-2, and THP-1 RNA. The colour scale is shown below, increased (orange) or decreased (blue) expression of the genes is reported for each sample. IOSE-hTERT_INT, the original SV40-immortalized normal ovarian surface epithelium cell line stably transfected with hTERT; LiCl, the IOSE-hTERT_INT cell line treated for 5 weeks with 5µM LiCl; FGF2/PDGF, the IOSE-hTERT_INT cell line treated for 5 weeks with 10ng/ml PDGF and 10ng/ml FGF2; RPMI_FGF2, the IOSE-hTERT_INT cell line treated for 5 weeks with 10ng/ml FGF2 maintained in RPMI (RPMI_FGF2); EOC, human epithelial ovarian cancer. **(B) Northern blot analyses.** Analyses of the transcript abundance of the unknown potentially coding clones. Shown are the results of clone 5000EIA09 (lane B) and 5000AFE01 (lane C), as expressed in the CaCo-2 cell line. GAPDH (lane A) was also checked as positive control.

Conserved sequence tags in the unknown clone sequences

The observation of conserved sequence tags (CST) in cross-species comparison may contribute to the identification of new genes or gene isoforms or regulatory motifs in the untranslated regions of mRNA.

A recently developed program – CSTminer (Mignone F et al, Nucleic Acids Res 2003; Castrignano T et al, Nucleic Acids Res 2004) - is able to identify statistically significant conserved tracts and to assess their coding or noncoding nature. To investigate further the 342 clones corresponding to potentially new genes I compared their sequences to the *Mus musculus* genome by using the CSTminer program. I observed at least one CST in 59 clones, and 12 clones contained coding CST (average coding potential score - CPS - 8.22). These 12 clones are very likely to contain a significant portion of a coding region, with 11 of them, falling in intergenic regions, the most reliable representatives of novel human genes. Forty-seven clone sequences containing a CST (11 with a coding CST) were also expressed in at least one tissue, while 33 clones (7 with a coding CST) were expressed in more than two tissues (Table 3).

A further confirmation of this hypothesis is given by the comparison of these 12 clones with the data contained in the DBTSS (Suzuki Y et al, Nucleic Acids Res 2004) collection: only 1 of them (clone 5000BHE02) matches with entries of this database and in particular with NM_024656 (*Homo sapiens* hypothetical protein FLJ22329, mRNA length 3593 bp) and the *M. musculus* counterpart NM_146211 (*M. musculus* hypothetical protein MGC38524, mRNA length 3231 bp). The 11 other clones do not match at all with known or predicted human or mouse genes or transcripts.

Finally, by considering also the 3' end and the internal sequences obtained it has been possible to confirm the potential coding capacity of these 12 clones; it seems, in fact, that all of the 12 clones of interest possess open reading frames (ORF) leading to the production of peptides, with the best result represented by clone 5000FJE08, which presents an ORF of at least 171 aa.

Northern blot analyses

The expression levels of 11 of the 12 unknown potentially coding clones have also been evaluated by Northern blot analyses on one cell line, CaCo-2. Only 2 clones, 5000EIA09 and 5000AFE01 (Fig.21B), are clearly expressed in the CaCo-2 cell line, while the 9 other clones give weak signals most probably indicative of low expression levels (data not shown).

Effects of the 12 unknown potentially coding clones on transfected cells

The biological effects of 8 of the 12 unknown potentially coding clones (data not shown), cloned in expression vectors, were tested by transfecting U2OS cells together with GFP under different experimental conditions.

Two of these clones, 5000BHE02 and 5000CBA02, do not seem to modify the overall aspect of the transfected cells as analyzed in the cells co-expressing the transfection marker GFP. On the other hand, six clones give rise to interesting phenotypes of different kinds. The most striking effect of clones 5000GCA01, 5000CIC06, and 5000GDH05 involves changes to the cytoskeleton (although they seem to act by following different mechanisms), while clones 5000AFE01 and 5000EDD03 induce changes at the level of the nucleoli structures. The most striking effect was shown by clone 5000FJE08. Fig.22 demonstrates that 24 h from the transfection the GFP marker seems to be concentrated in vesicular aggregates within the cytoplasm. The amount of such vesicles/aggregates increases during the next 24 h, leading to the most dramatic effect 48 h after transfection, when the cytoplasm is entirely filled with these vesicles and the cell nuclei have shrunk.

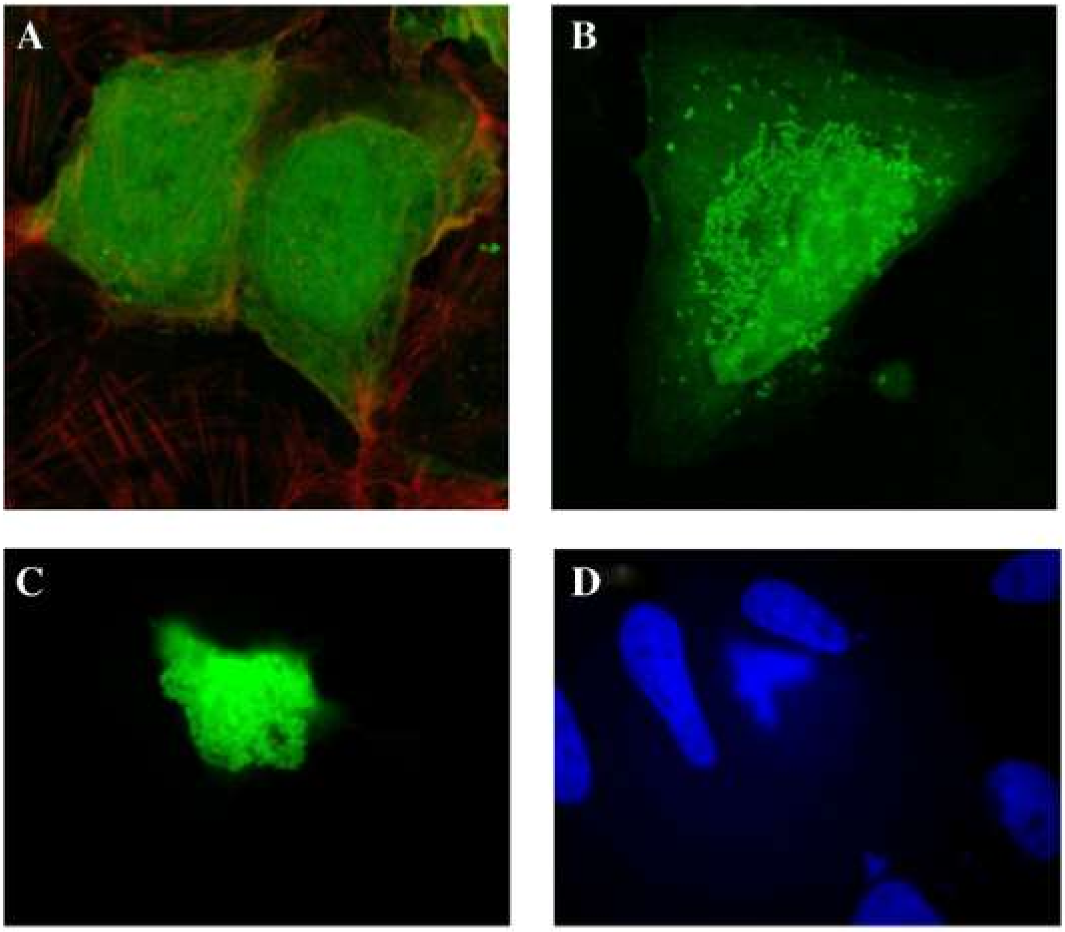


Figure 22. Effect of clone 5000FJE08 on U2OS transfected cells. (A) Negative control: cellular distribution of GFP. (B) 24 h after transfection. (C) Cytoplasm of a U2OS cell 48 h after transfection. (D) Apoptotic nucleus of a U2OS cell 48 h after transfection.

Full-length sequencing of cDNA clone 5000JE08

The complete sequence of clone 5000FJE08, responsible of the most striking biological effect among the 12 unknown protein-coding cDNAs, was recently obtained.

```
ATCCCCCCCCGGAAGGCGGCCTCGGCCCAGTGCACAGCGGGACCAGGCAGAGTTCGGGGAAAGCGTCGG
AGTTCGGGAGACCAGGGTCCAGCATGGGTTTCAGCACAGCAGACGGCGGGGGCGGCCCAGGCGCCCGGG
ATCTGGAATCTCTTGATGCCTGTATCCAGAGGACGCTCTCTGCCTTGTACCCACCGTTTGAAGCCACGGCA
GCCACGGTGCTCTGGCAGCTGTTCAGCGTGGCCGAGAGGTGCCACGGTGGGGACGGGCTGCACTGCCTCA
CCAGCTTCCTCCTCCCAGCCAAGAGGGCCCTGCAGCACCTGCAGCAGGAAGCCTGTGCCAGGTACAGGGG
TCTGGTCTTCCTGCACCCAGGCTGGCCGCTGTGCGCCCATGAGAAGGTGGTGGTGCAGCTGGCGTCCCTG
CACGGAGTCAGGCTCCAGCCCGGGGACTTCTACCTGCAGGTCACGTCGGCGGGGAAGCAGTCAGCTAGA
CTGGTCTTGAAATGCCTGTCCCGGCTGGGAAGAGGCACAGAGGAAGTCACCGTCCCTGAGGCCATGTATG
GCTGTGTCTTCACGGGGGCGTTCCTGGAGTGGGTGAACCGGGAGCGGCGCCATGTCCCCCTGCAAACCTG
CTTGCTGACCTCAGGCTTGGCCGTCCACCGAGCCCCGTGGAGCGACGTCACTGACCCTGTCTTTGTCCCCA
GCCCTGGAGCCATCCTGCAGAGCTACTCCAGCTGCACAGGGTCCTGAGCGGCTGCCCAGCAGCCCCTCAG
AGGCCCCAGTCCCCACCCAAGCCACAGCAGGCCCCCATTTCCAGGGAAGCGCCTCTTGCCCCGACACCCT
GACCTCACCCTGCCGCCGAGGGCATACGGGCAGCGACCAGCTCAGGCACCTTCCTTATCCAGAAAGAGCC
GAGCTGGGAAGCCCCAGGGACCCTGTCTGGAAGCTCAGACAGGGACTTCGAAAAGGTCAGCCCCTCAGA
GCAGGGCCCACGGATGCCCCCTGAGAACTGTGGGGGGTCGGGGGAGAGGCCGGACCCCATGGACCAGGA
GGACAGACCCAAGGCCCTCACCTTCCACACAGACCTGGGCATCCCGAGCAGCAGGAGGCGGCCGCCGGG
GGACCCCACTTGTGTGCAGCCTAGACGCTGGTTCAGGGAGTCGTACATGGAAGCCTTGCGGAACCCCATG
CCCCTGGGCAGCTCTGAGGAGGCCCTCGGGGACCTGGCCTGCAGCTCCCTGACTGGAGCCAGCAGGGACC
TGGGGACTGGGGCAGTAGCCAGTGGGACCCAGGAGGAAACCTCTGGCCCCCGGGGAGACCCCCAACAGA
CCCCAAGTCTAGAGAAGGAGAGGCACACACCCAGCCGGACAGGTCCAGGAGCTGCAGGGCGGACTCTTC
CCAGGGAGATCTCGGTCCTGGGAAAGGGCACCCAGAAGCTCCAGAGGGGCCCAGGCTGCAGCCTGCCAC
ACCTCCCACCACTCAGCAGGCTCCAGGCCTGGGGGCCCACCTAGGAGGACAAGCTGTGGGGACCCCAAA
CTGTGTCCCAGTAGAGGGTCCCGGCTGCACCAAAGAGGAAGACGTTCTTGCATCCTCAGCCTGTGTCAGC
ACAGACGGCGGCAGCCTCCATTGCCACAACCCCAGCGGGCCTTCCGATGTGCCTGCCCGGCAGCCACACC
CCGAGCAAGAAGGGTGGCCACCCGGCACAGGAGACTTCCCCAGCCAGGTGCCCAAGCAGGTGCTGGACG
TCAGTCAGGAGCTGCTGCAGTCCGGGGTCGTCACCCTCCCAGGGACCCGAGACCGTCATGGCAGAGCAGT
GGTGCAGGTCCGCACCAGGAGCCTGCTCTGGACCAGGGAACACTCGTCCTGTGCTGAGCTGACCCGCCTG
CTGCTGTACTTCCATAGCATCCCCAGGAAAGAGGTCCGGGACCTGGGGGCTGGTTGTCCTGGTGGATGCA
CGCAGGAGTCCAGCTGCCCCTGCCGTCTCCCAGGCCCTCTCAGGATTGCAGAACAACACATCTCCTATAA
TTCATCTATAAATTCATAGTATCTTGCTGTTGGTAGATAAAGAATCTGCATTTAGGCCTGACAAGGATGCA
ATAATTCAGTGTGAGGTCGTGAGCTCCCTGAAGGCCGTGCACAAATTTGTTGACAGCTGCCAGCTGACCG
CAGACCTCGACGGCTCCTTTCCCTACAGCCATGGTGACTGGATCTGCTTCCGTCAGAGGCTGGAACACTTC
GCTGCAAACTGTGAAGAAGCCATCATTTTCCTACAGAATTCATTCTGCTCACTGAACACCCACAGAACAC
CAAGAACAGCCCAGGAAGTCGCCGCTTTAATTGACCATCATGAGACGATGATGAAGCTTGTCCTGGAAGA
TCCACTGCTTGTGTCTTTCAGGCTGGAGGGGGGCACCGTCCTGGCGCTGCTGAGGAGAGAAGAGCTTGGC
ACAGAAGACAGCCGGGACACCTTGGAGGCCGCCACAAGCCTGTACGACCGAGTGGATGAGGAGGTGCAC
AGGCTGGTCCTCACCTCGAACAATCGTCTCCAGCAGCTGGAGCACCTCCGGGAGCTGGCGTCACTCCTGG
AAGGGAATGACCAGGTCAGAGCTGCAGGAGGAAGGCGGCCCGGTCAGCATCTCCTCCAGTCCAGGCTGG
CCAAGGCAACCCTCTCACCTTCACACTGTGTCTTTAGGGCCTTGATCTGATTTCCATTTGGAATGAAATTA
TGTTCAGGATAGGGCATGCCTTGCTGCTTGTGTAAAAAGAAATAAATTTTATTTTTTACGTGTGAGATACT
GATAAAAAAAAAAAAA
```

ATGpr (Salamov AA et al, Bioinformatics 1998), a program for identifying the initiation codons in cDNAs sequences, was used to study the coding potential of this clone. Two different ORFs were identified, thus confirming the initial results provided by CSTminer:

Reliability	Frame	Start (bp)	Stop (bp)	ORF length	Stop codon
0.63	2	992	1504	171	Yes
0.62	3	93	743	217	Yes

Further experiments will be soon performed to define which ORF is really used and what are the characteristics of the corresponding protein.

New BLAST queries were performed, too, to verify if the functional annotation could improve by using the complete sequence of the novel gene. First of all the EnsEMBL human genome database was queried, to define the exact location of the cDNA, and 11 matches were found with exons of 3 different unknown transcripts of the 5p15.33 region, suggesting that these different transcripts may be indeed produced by the transcription of the same gene.

Along that, the sequence of clone 5000FJE08 was also compared to the nucleotide and protein versions of the NCBI nr database, showing a partial similarity to the predicted sequence of "KIAA1909 protein, mRNA", which is associated to a *GTPase activator activity* but whose biological role and exact sequence remain hypothetical.

The presence of significant gaps in the latter BLAST results, and the fact that the function of the KIAA1909 protein still remains almost completely unknown, put great expectations on the results of the biological essays that will be soon performed.

PART 2.1 BIOINFORMATICS PLATFORM FOR THE IDENTIFICATION OF TRANSCRIPTION FACTOR BINDING SITES

The second part of the work presented in this thesis is focused on the analysis of the subset of novel genes previously identified and on their functional characterization, with cDNA microarrays used as a major source of information. In general, gene expression experiments give rise to gene clusters whose clear distinction between co-expression and co-regulation events becomes a key feature in predicting the molecular pathways that are involved in the basic cellular processes to be understood. Along this line, the predictive potential of existing algorithms for promoter discovery was then evaluated while working with human genome sequences, and their significant failure in performing analysis without *a priori* knowledge led to the planning, development, validation and use of a novel algorithm (called ScanPro) for the *de novo* identification of TFBS.

Identification of Transcription Factor Binding Sites

The first, most important application of motif search algorithms is their use in *gene expression experiments*, like cDNA microarrays experiments performed at LNCIB. After performing gene clustering, it is in fact essential to separate genes that are simply co-expressed by those which are really co-regulated, that is, to understand if a group of co-expressed genes shares a set of regulatory regions that can be the target of one or more Transcription Factors. Ideally, by analyzing the genomic neighbourhood of these genes, it should be possible to define the binding sites recognised by the common TF(s) and their locations in the genomic region. However, things are not this simple, from the very beginning.

A second possibility is to have a *genome-wide approach*, without considering gene expression results: all (or at least a large number of) the non-coding regions of an organism are examined, and overrepresented motifs can be suspected to play some role in the regulation of the genes, and therefore considered to be candidate TFBS.

A third very interesting application for these algorithms allows to *infer the function of an unknown gene* when homologous genes of known function are not available. If the TFBS of the unknown gene are identified and compared to those of known genes, it is possible to hypothesize that, if these binding sites are in common, known and unknown genes share the same function.

Finally, it is possible to extend the concept of motifs searching from the single binding site approach to a broader, *module-based approach*, where a module is defined as a group of at least 3-4 different TFBS that are conserved in different co-regulated genes and that maintain a constant overall distance with respect to one another. By encoding the single module as it has been previously done for single transcription sites, the algorithm can be easily converted to this new task.

As soon as I approached the topic of motif search, I had to face the same problem that already occurred during the sequence analysis phase: many tools were available, but none of them (from my point of view) was able to analyze the kind of dataset I was interested in (human genome sequences) efficiently extracting without *a priori* knowledge the information I was looking for, that is conserved sequences putatively corresponding to regulatory elements.

Particularly troublesome to all approaches, whatever their aim, was the presence of too much noise, whether it came from stretches of repeated sequences with no biological function or from the presence in the data set of more than one family of independent promoter motifs.

The algorithms for extracting conserved sets of single or combined words in a sequence are however growing increasingly more sophisticated and efficient, especially when using combinatorial approaches. Unfortunately, it is in the statistical evaluation of the motifs found that problems seem to persist, especially when errors are allowed and motifs may be composed of more than one element with adjacent parts standing at particular distances from one another.

For this reason I pushed the decision to develop a new algorithm for TFBS search, called ScanPro.

ScanPro algorithm

The problem of identifying a conserved (regulatory) element among a set of genomic regions can be formulated as follows: find a substring or a small similar subsequence that is common to many of the strings in the set. Let's consider DNA sequences over the *alphabet* $\Sigma = \{A,C,G, T\}$. I make use of the *Hamming distance* d_H to define the concept of "similarity" among substrings: the Hamming distance between two strings of the same length is the number of symbols that disagree.

Let's introduce the following notation:

- $|S|$ denotes the length of S;
- $S[i]$ is the i-th character of the string S;
- $S_{i,l} = S[i,\ldots,i + l -1]$ is the substring of l characters starting from $S[i]$;
- $S \lhd T$ denotes that S is a substring of T;

Let S_1, S_2 be two sequences of length m and $a_1, a_2 \in \Sigma$, then

$$d_H(a_1,a_2) = \begin{cases} 0, & \text{if } a_1 = a_2 \\ 1, & \text{if } a_1 \neq a_2 \end{cases} \quad , \quad d_H(S_1,S_2) = \sum_{i=1}^{m} d_H(S_1[i], S_2[i])$$

Karp - Rabin

The algorithm proposed by Karp and Rabin (Karp RM and Rabin MO, IBM J Res Dev 1987), solves the pattern discovery problem on the exact string matching background. This algorithm assumes that it is possible to efficiently shift a vector of bits and that it is possible to efficiently perform arithmetical operations on integers. To take advantage of this assumptions, a string can be seen like an integer, mapping each character of Σ in a digit using a function f_m. For example, in the DNA context, Σ = {A,C, G, T} can be mapped into Σ' = {0, 1, 2, 3}. It is now possible to define the following function:

$$H(S_{1,l}) = \sum_{i=1}^{l} |\Sigma|^{l-i} \cdot f_m(S[i])$$

$$H(S_{r,l}) = |\Sigma| \cdot (H(S_{r-1,l}) - |\Sigma|^{l-i} \cdot f_m(S[r-1]) + f_m(S[r+l-1])$$

It is possible to assert that there is an occurrence of a pattern P starting at position r of T if and only if $H(P) = H(T_{r,l})$.

Karp and Rabin introduced a method called the *randomized fingerprint* method, that preserves the spirit of the above numerical approach, but allows to deal with larger numbers in an extremely efficient way. It is a randomized method because it introduces a probability of error, but the probability that a false match occurs can be bounded.

ScanPro exploits only the first part of the algorithm idea, because while working with approximate string matching it is not simple to introduce an efficient hashing function.

Input: τ , l, d and q, where $\tau = \{T_1,\ldots, T_m\}$ is the set of strings (not necessarily of the same length), d is the number of errors allowed in comparisons and q is a parameter denoting the minimum size of the set of τ -elements containing a common substring of length l and with d errors. Without loss of generality, input strings of the same length n are considered.

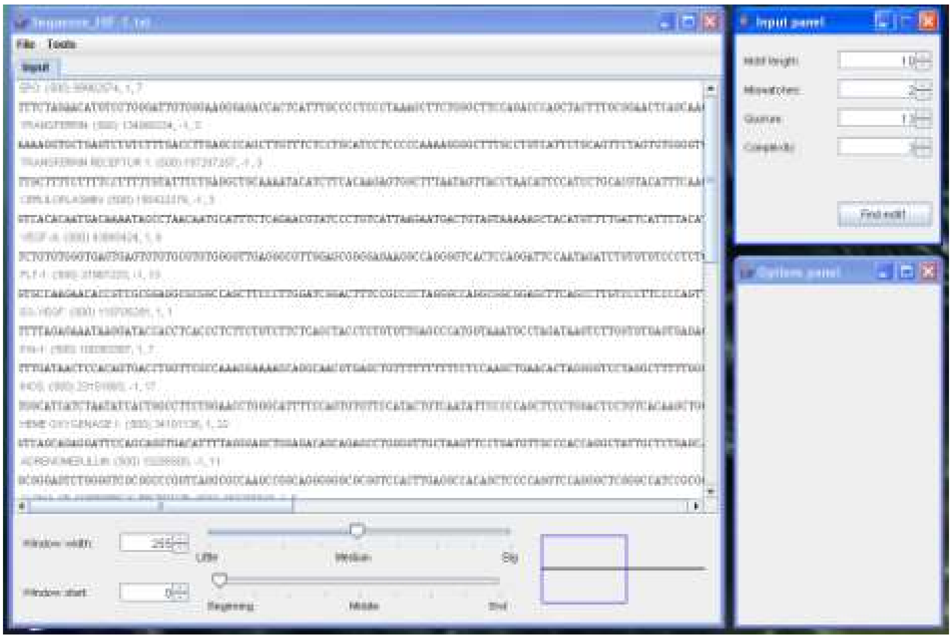
The (l, d, q) – *consensus problem* over τ is S iff

- $\exists$ a pattern p (*consensus*), |p| = l, such that $S \subseteq \{s \in \Sigma^{l} \mid d_H(s, p) \leq d\}$
- $\forall\ s \in S,\ \exists\ T_i \in \tau\ \mid s \lhd T_i$
- $|\{i \mid \exists s \in S, s \lhd T_i\}| \geq q$

A useful assumption is that, given a motif (TFBS), only some of its characters (nucleotides) are important for the binding of the Transcription Factor. This can happen as, along with the intrinsic degeneracy of DNA and the possibility to have experimental errors in the genomic sequences, a TF may be able to bind a subregion of the complete TFBS and still accomplish its function. This feature, that is considered also by statistical approaches when creating the weight matrices, allows to co-cluster different results, apparently corresponding to different motifs, simplifying and reducing the number of consensus TFBS generated by the algorithm. Thus, an approximated algorithm based on the concept of localized nucleotide mutation is proposed: a protein can “accept” one or more mutations in a binding site, but always in the same positions.

The *fixed-layout* (l, d, q) - *consensus problem* over τ is S_{fl} iff

- $S_{fl} \subseteq \Sigma^{l}$
- $\forall\ s' \in S_{fl},\ \exists\ T_i \in \tau\ \mid s' \lhd T_i$
- $|\{i \mid \exists\, s' \in S_{fl},\ s' \lhd T_i\}| \geq q$
- $\exists$ a set of indexes $fl = \{i_1, \ldots, i_d\}$, $1 \leq i_1 < \ldots < i_d \leq l$,
  such that for all s_i' and s_j' in S_{fl}, $s_i'[k] \neq s_j[k] \Rightarrow k \in fl$.



Motif length 10

Mismatches 2

Quorum 13

Complexity 3

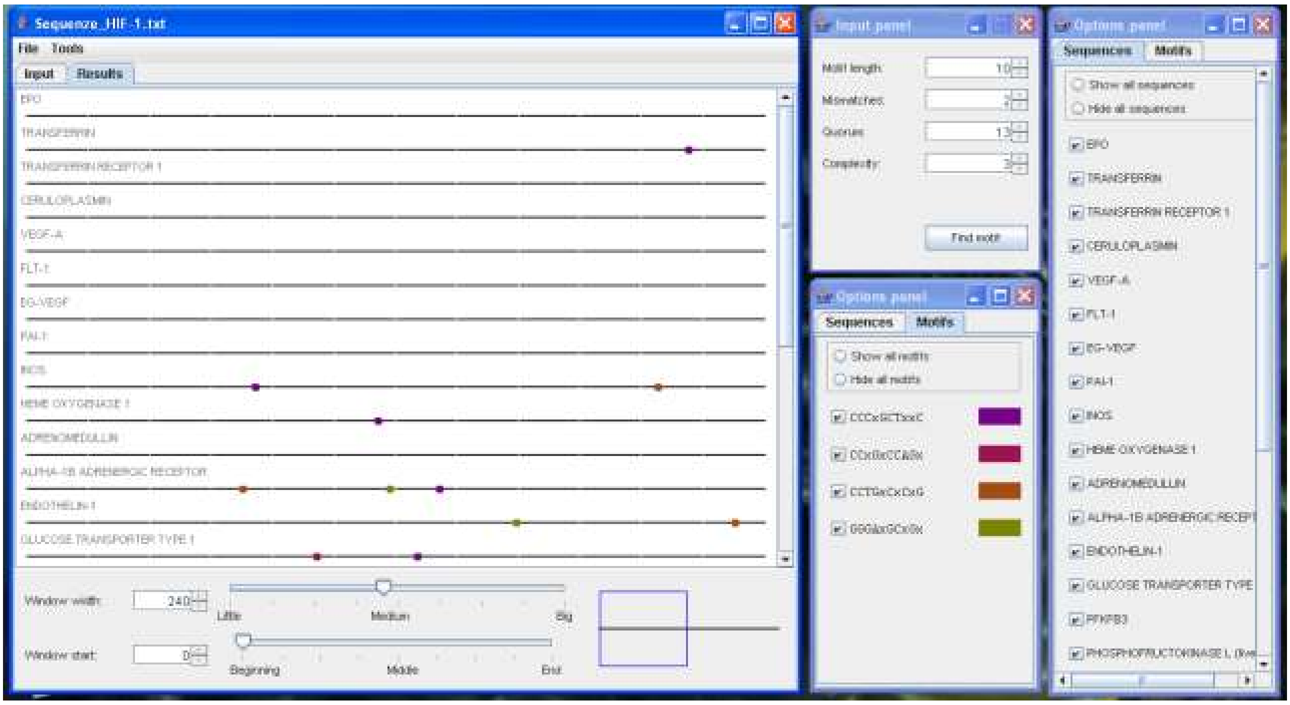
Window width 255

Window start 0

Figure 23. ScanPro graphic interface: input. The left window displays the input sequences; the Input panel allows to select the parameters that will be used to run the TFBS search.

Output: all the positions of the common subsequences of length l with d errors founded in at least q sequences, with the addition of the constraint that the errors, if occur, are in the same position (called layout).

This algorithm (presented at Workshop on Constraints Based Methods for Bioinformatics 2005, manuscript in preparation; see also *Personal Publications*) has the purpose of solving the *fixed-layout* (l, d, q) - *consensus problem*, and then extend the results to obtain an approximate solution for the (l, d, q) - *consensus problem*. The difference with solving immediately the former problem is in the generation of the consensus sequence and in the final complexity. The fixed-layout problem is a reduction of the consensus problem, with the aim to find some useful biological information.



Window width: 240

Window start: 0

Motif length: 10

Mismatches: 3

Quorum: 13

Complexity: 3

☑ CCCxSCTxxC 🟪

☑ CCxGxCCAGx 🟥

☑ CCTGxCxCxG 🟧

☑ GGGAxGCxGx 🟩

Figure 24. ScanPro graphic interface: output. The left window displays the location of the results on the input sequences; the Options panel shows the TFBS sequences and allows to remove some results, or some input sequences, from the visualization in the left window.

PART 2.2 EXPERIMENTAL TESTING

Selection of the best testing benchmark

In order to test ScanPro efficacy it was necessary to identify a benchmark to perform TFBS search through a combinatorial approach. For this reason two very interesting benchmarks (Pevzner PA and Sze SH, Proc Int Conf Intell Syst Mol Biol 2000; Tompa M et al, Nat Biotechnol 2005), that had been previously used for the assessment of publicly available programs for TFBS search, could not be used due to incompatibilities in the underlying theory.

I also wanted to validate ScanPro by immediately analysing the ultimate sequences that would have become its target, that is mammalian (and in particular human) sequences, so this was another reason for searching a more appropriate testing ground.

Many interesting datasets, often associated with exhaustive gene expression results, existed and could have been used in this process. However, in order to obtain an output useful for an easy initial debugging phase, the benchmark had to be quite simple.

The computational time required by ScanPro, for instance, had never been verified on nucleotide sequences and the analysis of a list made by hundreds of genes could have required too much time. The selection of a subset of sequences could have been a solution, but the risk of loosing precious data was too high, as TFBS searches are performed by ScanPro without *a priori* knowledge.

Two other critical features were the length of the expected TFBS and the number of mutations introduced, as the lack of statistical considerations during motifs search could have led on one hand to the absence of any conserved element, if its length or the number of mutations were too high and if the differences between the examined sequences were significant, and on the other hand could have produced consensus sequences almost completely different from the expected TFBS in the case of an exaggerated number on mutations.

A final consideration was however done in order to simplify the analysis. If a blind search had to be done and no previous knowledge was going to be used, it should have been useful to consider that, in general, Transcription Factors that regulate the expression of a group of genes involved in a given biological process tend to bind genes promoters in the same region (Segal E et al, Nat Genet 2005), thus reducing the dimension of the sequences taken into account.

Given these premises, my choice fell on a group of genes whose expression was regulated by the Hypoxia Inducible Factor 1 transcription factor (Semenza GL and Wang GL, Mol Cell Biol 2002; Wenger RH, FASEB J 2002). This factor, that was known to lead to the transcriptional induction of the gene encoding erythropoietin in conditions of oxygen deprivation, was recently found to regulate the expression of many more genes, and due to this discovery the molecular principles of oxygen sensing (Fig.25) are currently being elucidated.

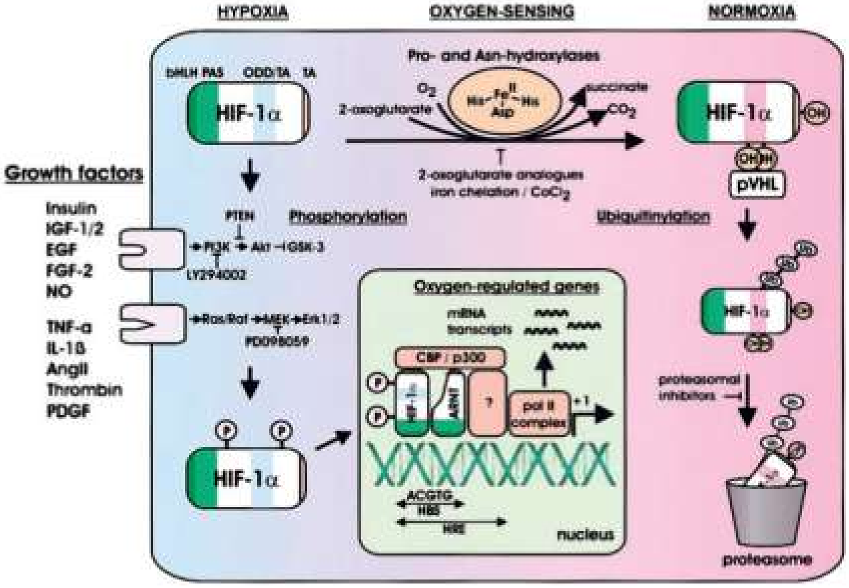
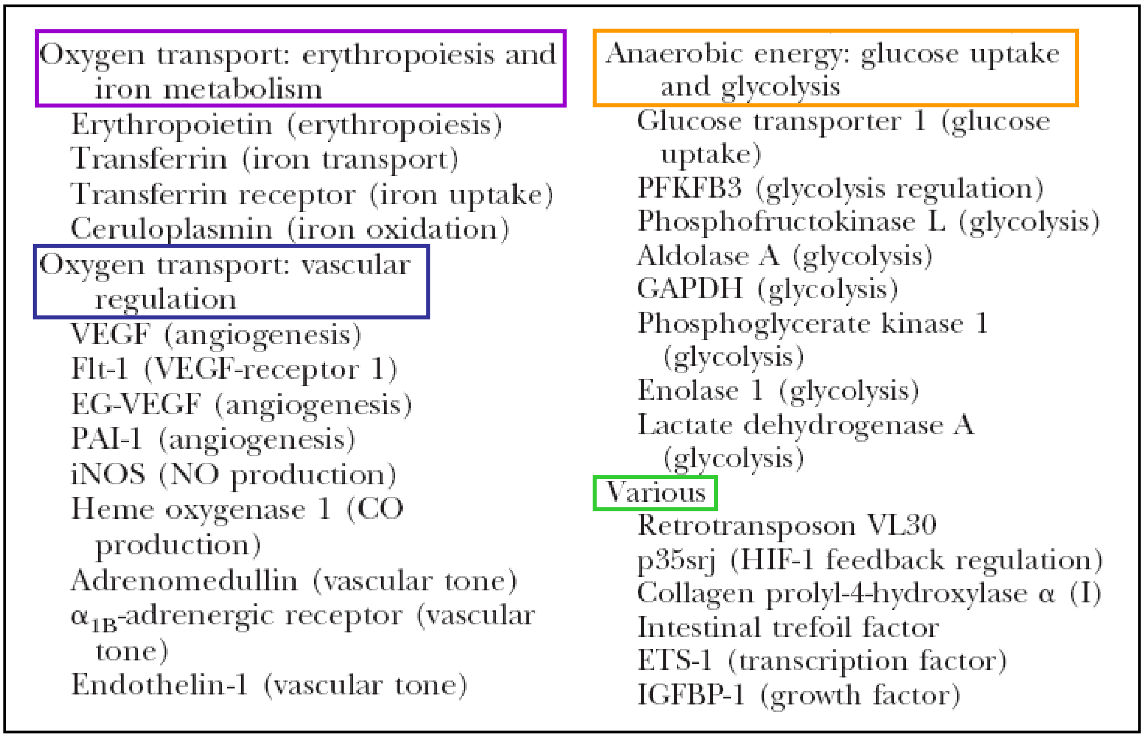


Figure 25. Simplified schematic model of oxygen sensing and HIF regulation (From: *Wenger RH, FASEB J 2002*)**.**

HIF-1 is a phosphorylation-dependent and redox-sensitive protein composed of two different subunits: HIF-1 α is a novel protein, while HIF-1β is identical to the previously identified heterodimerization partner of the dioxin receptor/aryl hydrocarbon receptor (AhR), called AhR nuclear translocator (ARNT). Under hypoxic conditions, HIF-1α is stabilized, undergoes modifications, translocates into the nucleus, recruits cofactors, and activates gene expression. Under normoxic conditions, the ODD domain of HIF-1α is hydroxylated by oxygen-dependent prolyl hydroxylases (targeting HIF-1α for pVHL-mediated proteolytic destruction) and the carboxyl-terminal TA domain is hydroxylated by an oxygen-dependent asparaginyl hydroxylase (blocking the interaction with the CBP/p300 coactivator). Many processes involved in oxygen homeostasis are mediated by hypoxia-inducible factors (HIFs), which transcriptionally regulate the expression of ~ two dozens of target genes (Table 2). This list of regulated genes includes genes involved in iron metabolism and transport (as iron is required for heme formation), modulation of vascular tone and density, and in glucose metabolism (as anaerobic glycolysis becomes predominant when oxygen supply is limited). Thus, HIFs represent the link between oxygen sensors and effectors at the cellular, local, and systemic level.

This dataset was chosen for two reasons in particular: first of all, the limited (26) number of genes that would have been analyzed; then, because the HIF-1 binding site is short and very simple:

$$\{A \mid G\}CGTGA$$



Oxygen transport: erythropoiesis and iron metabolism

Erythropoietin (erythropoiesis)

Transferrin (iron transport)

Transferrin receptor (iron uptake)

Ceruloplasmin (iron oxidation)

Oxygen transport: vascular regulation

VEGF (angiogenesis)

Flt-1 (VEGF-receptor 1)

EG-VEGF (angiogenesis)

PAI-1 (angiogenesis)

iNOS (NO production)

Heme oxygenase 1 (CO production)

Adrenomedullin (vascular tone)

α_{1B}-adrenergic receptor (vascular tone)

Endothelin-1 (vascular tone)

Anaerobic energy: glucose uptake and glycolysis

Glucose transporter 1 (glucose uptake)

PFKFB3 (glycolysis regulation)

Phosphofructokinase L (glycolysis)

Aldolase A (glycolysis)

GAPDH (glycolysis)

Phosphoglycerate kinase 1 (glycolysis)

Enolase 1 (glycolysis)

Lactate dehydrogenase A (glycolysis)

Various

Retrotransposon VL30

p35srj (HIF-1 feedback regulation)

Collagen prolyl-4-hydroxylase α (I)

Intestinal trefoil factor

ETS-1 (transcription factor)

IGFBP-1 (growth factor)

Table 2. HIF-1 target genes. Only proven (by DNA binding or transcription assays) direct HIF-1 target genes are listed.

The genomic regions to be analysed were extracted from the “homo_sapiens_core_26_35” MySql database downloaded from the “Download” page of the EnsEMBL Genome Browser (http://www.ensembl.org/info/data/download.html). In order to limit the complexity of the results, I decided to initially focus my attention on the (-500, -1) region of each gene, with respect to the Transcription Start Site as defined by EnsEMBL. The corresponding sequences were extracted by using the Application Programme Interfaces provided with EnsEMBL.

Validation of the testing benchmark: analysis with publicly available TFBS search tools

First of all I decided to use some of the publicly available programs for TFBS search to analyse the HIF-1 dataset, as HIF-1 TFBS was probably only one of the binding sites located into the considered genomic regions.

Transfac. Transfac (Wingender E et al, Nucleic Acids Res 2001) is a database on eukaryotic cis-acting regulatory DNA elements and trans-acting factors, covering the whole range from yeast to human. The database is made of different tables:

SITE gives information on (regulatory) transcription factor binding sites within eukaryotic genes. **GENE** gives a short explanation of the gene where a site (or group of sites) belongs to. **FACTOR** describes the proteins binding to these sites, while the **MATRIX** table gives nucleotide distribution matrices for the binding sites of transcription factors.

The Transfac Factor table reports a list of genes that hold a HIF-1 binding site, as follows (Table 3).

Target	Location	Start	End	Motif
Aldolase A	Exon H	-204	-180	gtggtccgaGTCACGTCcgaggggg
Enolase 1	TSS	-416	-397	agggccgGACGTGgggcccc
	TSS	-368	-349	ggagTACGTGACggagcccc
	TSS	-390	-371	acgctgagTGCGTGCGggac
Erythropoietin	3' enhancer	3454	3471	gcccTACGTGCTgtctca
Endothelin-1	TSS	-132	-113	ctccggctGCACGTtgcctg
Transferrin	-3.6 kb	170	201	ttccTGCACGTAcacacaaagCGCACGTAtttc
Transferrin receptor	TSS	-97	-77	cgcgagcgTACGTGCCtcagg
VEGF	TSS	-982	-963	cagtgcaTACGTGggctcca

Table 3. HIF-1 target genes as reported by the Transfac Factor Table (Accession Number T01609). For each gene are reported the location, with respect to the Transcription Start Site, the start and end position, and the identified motif.

As can be clearly seen, none of the reported TFBS exactly corresponds to the expected motifs. The reasons of this behaviour could be due to the following reasons: the first one is biological, and is based on the fact that not all the nucleotides of a TFBS have the same importance in the TF-TFBS interaction; the second one is based on the algorithm used for TFBS search, which makes use of weight matrices in the case of Transfac.

Along with the reported results, some other sites were reported, too, but no information was available and hence they are not presented.

The next step consisted in examining the existing TFBS in each one of the 26 genes supposed to be regulated by HIF-1, by looking at the information contained in the Transfac Gene table (Table 4).

Gene	Motif	Start	End	Function	Binding Factor
Erythropoietin	CCCCTGGCTCTGTCCCACTCCTGG	3380	3411		
	GCAGCAGTGCAGCAGGTCCAGGTCC	3403	3427		
	GTCCGGG	3424	3430		
	AAACGAGGGGTGGAGGGGG	3431	3449		
	CTGACCTCTCGACCTACCGGCCTA	3482	3504	3' enhancer	Tf-LF1 (T01176) Tf-LF2 (T01177)
Transferrin	GAGGGCGGGAAGTTTTCCAGCCCA	-614	-591		
	TCTTTGACCTTGAGCCCAGCT	-474	-454		Tf-LF2 (T00827) Tf-LF2 (T01229)
	TCCTCCCCCAAAAGGG	-440	-425		
	CTGTGCTGGACTCCTTCCACTCGCGGGTCGTC	-193	-162		CTF (T00174) NF-1 (T00535) (both CCAAT-binding factors)
	GGGCGATTGGGCAACCCGGC	-103	-83		C/EBPalpha (T00105)
	AACACGGGAGGTCAAAGATTGCGCCC	-76	-48		COUP-TF1 (T00149) HNF-4alpha1 (T00372) Tf-LF1 (T00825) Tf-LF1 (T00826) HNF-4alpha2 (T02422) HNF-4alpha1 (T02429)
	TGGAATAAAG	-34	-18	TATA box	TFIID (T01175)
	CTCTTTGTTTGCTTTGCTTCTGTGTCAACTGGGCAACATTTGGAAACAACAAATATTGGTTCAG	58	125		
	GTTTGCTTT	64	72		EBP40 (T01106) EBP45 (T01107)
	CTGGTCAG	145	161		
	TTAAGGGCCAACTCGTG	189	222		
	GTTTTGCCCTGTTTGCGGTGAAAGATTTGGCTCATGCTTGGGTTGGT	232	281		
Transferrin Receptor 1	CAGGAAGTGACGCACAGC	-80	-63		TREF1 (T01069) TREF2 (T01070)
Ceruloplasmin					
VEGF					
FLT-1					
PROK1					
Serpine 1	CCAGTGAGTGGGTGGGGCTGGAACATG	-86	-60		HLTF (T04146) HLTF (Met123) (T04147)
NOS2A	TAAATAAATAAAT	-742	-730		FOXF1 (T02461)

Gene	Motif	Start	End	Function	Binding Factor
Heme Oxygenase 1					
Adrenomedullin					
Alpha-1B Adrenergic Receptor					
Endothelin-1	TggccTTATCTccgg	-141	-127		GATA-2 (T00308)
	GTGACTAA	-109	-102		AP-1 (T00029) c-Fos (T00123) c-Jun (T00133)
SLC2A1					
PFKFB3					
Phospho-fructokinase L					
Aldolase A	CGTTCCGCCCTCCCCCATTGCCAACATTCTGGCTGAGTCACGGCGCCCCAG	-311	-261		AP-1 (T00029) AP-1 (T00032) Sp1 (T00752)
	GTCCGAGGGGGGTGGGGAGGGATCGT	-190	-165	enhancer for N and H promoter	
	GGCGCCCGCCCCTTCCTAGC	-158	-140	enhancer for N and H promoter	
	CTGGGCTGGCCTCTCGGGGGCGGCCCGT	-131	-104		Sp1 (T00759)
GAPDH	CCCGCCTCtcagCCTTTGAAagaaagaaagggg	-480	-435		SRY (T00996) IRE-ABP (T00998) SRY (T01986) SRY (T01987)
Phosphogly-cerate kinase 1					
Enolase 1					
Lactate Dehydrogenase A					
CITED2					
P4HA1					
Intestinal Trefoil Factor					
ETS-1					
IGFBP-1	cactagCAAAACAaactTATTTTGaacac	-124	-96		HNF-3alpha (T00371) HNF-3B (T01049) HMG I (T01851) HMG Y (T01980) FOXO3a (T02938)
	TGCGGCGCTGCCAATCATTAAC	-79	-53		HNF-1A (T00368) HNF-1B (T01950) HNF-1C (T01951)

Table 4. TFBS located in the (-500, -1) region of the 26 HIF-1 regulated genes - Transfac Gene Table data. For each gene are reported the motif, the start and end position, the function of the element (if any) and the identified binding factor.

All the reported TFBS correspond to binding sites for TFs with a quite broad action range, and none of them correspond to the expected HIF-1 binding site. Almost half of the 26 HIF-1 regulated genes do not present any significant TFBS reported in Transfac, while for some others there is no information available about any bound protein.

From the available data it is however possible to identify a subgroup of genes that are involved in inflammatory response and cell proliferation control (Table 5).

Gene	Binding Factor	Function
Transferrin	C/EBPalpha (T00105)	inhibits cell proliferation by enhancing the level of p21 (WAF-1)
	COUP-TF1 (T00149)	represses RXR-mediated activation and retinoic acid responses; also antagonizes HNF-4-mediated activation
	HNF-4alpha1 (T00372)	transcriptional activator, may be antagonized by ARP-1 and/or COUP-TF; HNF-4alpha is involved in regulating cancer cell transmigration by modulating the Fas-FasL system
Endothelin-1	GATA-2 (T00308)	functional cross talk between RA and GATA-2-dependent pathways
	AP-1 (T00029)	down-modulated by glucocorticoids through direct interaction with GR; induced by TPA
Aldolase A	AP-1 (T00029)	down-modulated by glucocorticoids through direct interaction with GR; induced by TPA
	Sp1 (T00759)	highly specific cooperation with NF-kappaB; may play a role in regulation of IL-1alpha gene expression
IGFBP-1	HNF-3alpha (T00371)	inhibits HNF-1 action on aldolaseB promoter; RA induces transcription; probably involved in some steroid hormone regulatory circuits
	FOXO3a (T02938)	may have regulatory roles in immune response; triggers apoptosis by inducing the expression of genes that are critical for cell death, such as the Fas ligand gene and the Bcl-2 interacting mediator Bim
	HNF-1A (T00368)	synergistic interaction with C/EBPalpha

Table 5. Involvement of HIF-1 regulated genes in inflammatory response and cell proliferation control. For each gene are reported the binding factor and the associated function that suggest an involvement of HIF-1 regulated genes in these two biological processes.

YMF. Considering the results obtained with Transfac as only partially satisfying, I decided to analyze the dataset with YMF (Sinha S and Tompa M, Nucleic Acids Res 2003). YMF is a program that detects statistically overrepresented words (motifs) in DNA sequence, with "motifs" corresponding to short string of nucleotides, degenerate symbols, and spacers. It is possible to specify the characteristics of the motifs to be detected: **'Motif size'** is the number of non-spacer characters in a motif. **Spacers** ('N's) are constrained to be in the centre of the motif. **Degenerate symbols** allowed in a motif are **R** (purine - A or G), **Y** (pyrimidine - C or T), **W** (A or T), and **S** (C or G). YMF constructs a third-order Markov model of the background sequences (all known promoter sequences) for the organism under study, *Homo sapiens* in this case. Given a set of sequences, YMF does an enumerative search among all motifs that match the specified characteristics, scoring each motif for its significance, and outputs the top several motifs, sorted by their significance.

Given the HIF-1 binding site, I selected "**6**" as the "**Motif size**", with a maximum of "**1**" tolerated "**degenerate symbol**" and **no spacers** in the middle of the results.

The five best results that were obtained are presented in Table 6.

They were compared to the Transfac Matrix table entries in order to link them to some known TFBS.

Motif	Count	z-score	Annotation
ACACAC	42	16.68	modulator recognition factor 2 (M00454)
ATARAT	22	6.73	SOX-9 (M00410) GATA-6 (M00462) TATA binding protein (M00471) Pax-2 (M00486)
GCCCCS	88	6.28	STAT5A (M00460)
ACGTRC	17	5.23	AREB6 (M00412)
AGCGAR	24	4.65	IRF-7 (M00453)

Table 6. YMF identification of the five best conserved motifs. When degenerate symbols were present, all the combinations were tested to identify associated TFs annotation.

These results identify TFBS that are present in quite great quantity in the (-500, -1) region, probably indicating a low specificity in their functional mechanism. Along this, in many cases the identified hexamers constitute only a fraction of the whole TFBS recovered in Transfac, suggesting that they could simply represent short sequences (sometimes with low complexity) interspersed within the examined regions, thus being devoid of any real regulatory function. This seems also confirmed by the fact that it is quite difficult to identify a common pathway that may involve two or more of the TFs binding to the conserved motifs. STAT5A and AREB6, binding respectively to the third and

fourth conserved motifs, are indeed involved in immune response, in particular in IL2 response, but associating the conservation of these two motifs to this biological function is purely speculative.
The fourth result is also very similar to the expected HIF-1 TFBS. However, YMF identifies it as putatively being ACGTG**C**, instead of ACGTG**A,** and its annotation is **"**AREB6 binding site". This is due, as stated above, to the fact that this hexamer is located in the middle of the AREB6 binding site (A**ACGTAC**CTGTGA) and the "mutation" that would be required to convert it into the latter motif would occur in a critical region for the DNA-protein interaction.
Finally, it is interesting to reaffirm how the YMF-generated results have almost nothing to share with those previously obtained from Transfac, thus demonstrating the (excessive) difference among the results obtained from different TSDB search algorithms. The only "conserved" element, that is however present in only two different genomic regions, is luckily the HIF-1 binding site corresponding to the fourth motif identified by YMF: this element is in fact located within the HIF-1 binding sites of the Erythropoietin and of the Transferrin Receptor as reported by the Transfac Factor Table, although its last nucleotide is different from the theoretically expected one.

Consite. The third approach I considered was Consite (Lenhard B et al, J Biol 2003), which integrates binding site prediction generated with high-quality transcription factor models and cross-species comparison filtering (phylogenetic footprinting). This tool is generally used to perform TFBS queries by comparing sequences from two different organisms, looking for evolutionary conserved regions that may contain regulatory elements. However, Consite also contains a collection of metazoan transcription-factor-binding profiles that is used as a reference database for a "single genome" approach by BLASTing the genomic sequence of interest. Unfortunately, HIF-1 is not among the many TFs stored into this database and for this reason the analysis could not take place.

CSHLmpd. Finally, I decided to use a recently developed database, called Cold Spring Harbor Laboratory Mammalian Promoter Database (Xuan Z et al, Genome Biol 2005), that was built by exploiting both the conservation among mammalian genomes and an *ab initio* promoter prediction. This tool allowed me to identify the first exons of each one of the 26 studied genes, along with the known or predicted promoters associated to these regions. Unfortunately, only 9 out of 26 genes were associated to annotated elements (Table 7), with four of these being poorly meaningful, as containing the broadly distributed SP1 binding site, so the search for HIF-1 binding sites was almost impossible to complete. The remaining clones, in fact, did not give significant results, being completely devoid of any information or, sometimes, being only associated to an unknown sequence with a putative regulatory function. Only one gene, Heme Oxygenase 1, was associated with the conserved element ctggcccACGTGAcccgc containing one of the two expected forms of the HIF-1 binding site. None of the five YMF results were identified, as well as almost none of the results obtained from Transfac. Only in the case of the Transferrin gene five matches were obtained, corresponding to the positions (-474, -454), (-440, -425), (-193, -162), (-103, -83) and (-76, -48), this being due to the presence of six Transfac entries in the CSHLmpd Transferrin table.

Gene	Known Motif	Conserved unknown sequence
Transferrin	consensus Sp1 binding site	TCTTTGACCTTGAGCCCAGCT (DR I) CTGTGCTGGACTCCTTCCACTCGCGGGTCGTC (CR) GGGCGATTGGGCAACCCGGC (PR II) AACACGGGAGGTCAAAGATTGCGCCC (PR I) TCCTCCCCCAAAAGGG (DR 0)
Transferrin Receptor	consensus Sp1 binding site	
Serpine 1	pot. NF1 binding site; poly(AC) stretch; glucorticoid regulatory element	
Heme Oxygenase 1	newly defined heat-shock element	ctggccCACGTGACccgc (HIF-1 binding site) ccagactttgtttcccaaggGTCATATGACtgctcctctccaccccacactggc
SLC2A1		agccaatggccggggtcctataaacgctacggtccgcgcgctctctggcAAGAGGCAAGA
Phospho-fructokinase L		cgcgcgcggggcggggcggggacggcgacgcggcgcaggcggcgggagtGCGAGCTGGGC
GAPDH	IRE-A, insulin-response element A	
Phosphoglycerate kinase 1	GC-box; CCAAT-box	CGGTTCGCGGCGTGCCGGACG (D); AGACGGACAGCGCCAGGGAGCAATGGCAGCGCGCCGAC (D)
Enolase 1	HBS C; HBS D; HBS U	
Lactate Dehydrogenase A		acgtcagcatagctgttccacttaaggcccctcccgcgcccagctcagaGTGCTGCAGCC
Endothelin-1	AP1 consensus sequence; SP1 consensus sequence	
IGFBP-1	SP1 consensus sequence	

Table 7. CSHLmpd results. For each HIF-1 regulated gene matching with entries of the database are reported the associated known motifs (9 entries) and/or the conserved unknown sequences (6 entries). 3 genes present both annotations.

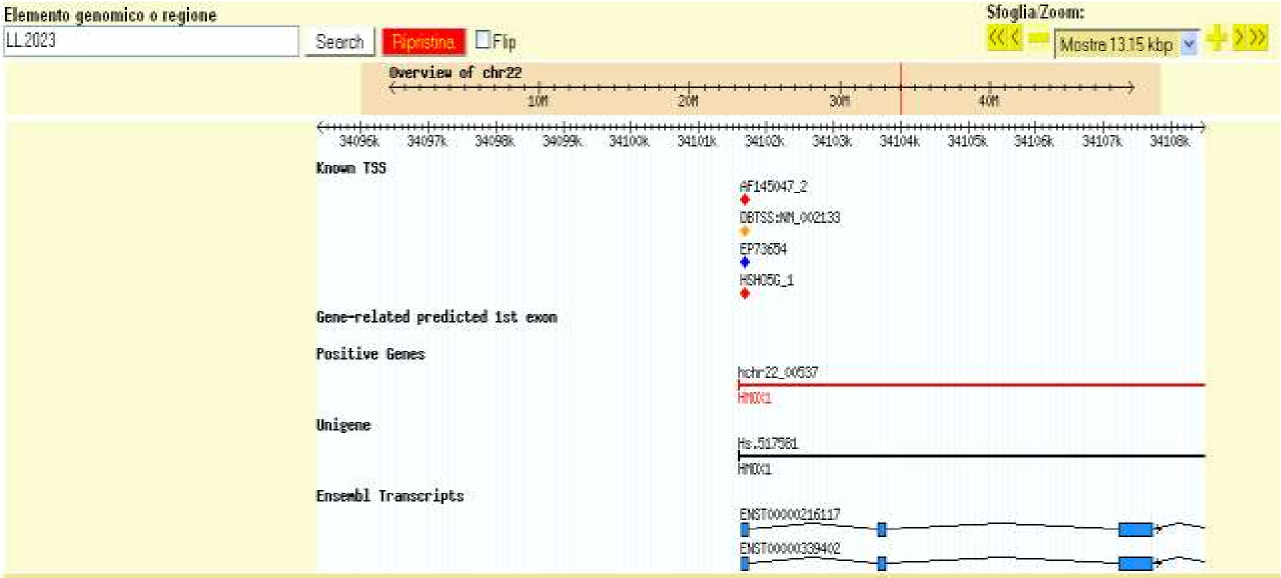


Figure 26. CSHLmpd results for Heme Oxygenase 1 (Gene ID 2023). This is the only region were it was possible to locate an HIF-1 binding site. Element HSHO5G_1, on the other hand, contains a "newly defined heat shock element".

Validation of the testing benchmark: analysis with ScanPro

After analyzing the (-500, -1) regions of the 26 HIF-1 regulated genes with the three selected public tools I finally applied ScanPro to the same dataset.

As previously described, ScanPro looks for conserved elements with a fixed Motif size and a fixed number of tolerated "mutations". Due to the structure of the analysed TFBS, the **Motif size** was fixed to "**6**" and the number of **mutations** to "**1**", as the HIF-1 binding site itself has this length and this degree of variability.

The first step of the ScanPro analysis led to the generation of a list of conserved elements containing one mutation located in one of the six available positions of the resulting hexamers. The following step consisted in generating a consensus motif where the mutations were replaced by a single defined character, if one of the four nucleotides was over-represented in that given position, or by one of the IUPAC symbols currently used for defining ambiguities (**R** (purine - A or G), **Y** (pyrimidine - C or T) and **N** (undefined)).

The exact results, that is the presence of one of the two expected HIF-1 binding sites in the analyzed regions, are depicted in Table 8.

Gene	Motif	Consensus	Position
Serpine 1	xCGTGA	ACGTGA	-454
	ACxTGA	ACGTGA	-131
Heme Oxygenase 1	xCGTGA	ACGTGA	-40
Phosphoglycerate kinase 1	xCGTGA	ACGTGA	-226
Enolase 1	ACxTGA	ACGTGA	-496
Adrenomedullin	ACxTGA	ACGTGA	-443
Phosphofructokinase L	xCGTGA	GCGTGA	-158

Table 8. ScanPro analysis: exact results. Six genes were identified presenting one of the two expected HIF-1 binding site. The motif, as well as the derived consensus sequence, is presented as well as its position with respect to the Transcription Start Site.

These initial results were encouraging, indeed, as for the first time TFBS prediction succeeded in identifying six exact binding sites by examining the HIF-1 dataset.

However, the output produced by ScanPro revealed other very interesting features, like shown in Table 9.

Gene	Motif	Consensus	Position
Erythropoietin	TCACGx	TCACGy	-211
	TCACGx	TCACGy	-135
Phosphoglycerate kinase 1	TCACGx	TCACGy	-342
Enolase 1	TCACGx	TCACGy	-75
P4HA1	TCACGx	TCACGy	-87

Table 9. ScanPro analysis: reverse complement exact results. Four genes were identified presenting one of the two expected HIF-1 binding site on the reverse complement sequence of the examined genomic region. The motif, as well as the derived consensus sequence, is presented as well as its position with respect to the Transcription Start Site.

It seems that some of the HIF-1 regulated genes (including Erythropoietin, the first identified HIF-1 target) do not bind the transcription factor on the same DNA strand that is supposed to be transcribed but on the other one, as shown by the conserved consensus TCACGy, exactly the reverse complement of the expected ACGTGA site. There seems to be no evidence, in literature, of this kind of "enhancing" effect of HIF-1 on its target genes, but the presence of this feature in five different regions of four known HIF-1 regulated genes, along with the proved efficacy of ScanPro in identifying regulatory elements, suggests to avoid the simple removal of these results as "false positive", as this could represent an alternative regulatory method used by HIF-1.

Finally, I obtained a huge amount of data that, although not containing the exact expected TFBS, greatly strengthened the predictive ability of ScanPro, as the output was however more similar (and quite often nearly identical) to the exact binding sites than almost the majority of the "positive" results obtained thus far with other predictive tools, including Transfac. Table 10 presents the results containing one "mistake" with respect to the exact HIF-1 binding sites.

Gene	Motif	Consensus	Position
Transferrin	AxGTGA	AGGTGA	-143
Transferrin Receptor	AxGTGA	AGGTGA	-91
NOS2A	AxGTGA	AGGTGA	-395
	AxGTGA	AGGTGA	-328
Heme Oxygenase 1	AxGTGA	AGGTGA	-481
	AxGTGA	AGGTGA	-20
PFKFB3	AxGTGA	AGGTGA	-312
Phosphofructokinase L	AxGTGA	AGGTGA	-472
Enolase 1	AxGTGA	AGGTGA	-357
	AxGTGA	AGGTGA	-86
CITED2	AxGTGA	AGGTGA	-412
P4HA1	AxGTGA	AGGTGA	-486

Gene	Motif	Consensus	Position
VEGF	GxGTGA	GAGTGA	-493
	GxGTGA	GAGTGA	-489
PROK1	GxGTGA	GAGTGA	-410
	GxGTGA	GAGTGA	-406
	GxGTGA	GAGTGA	-257
NOS2A	GxGTGA	GAGTGA	-214
	GxGTGA	GAGTGA	-194
	GxGTGA	GAGTGA	-37
Alpha-1B Adrenergic Receptor	GxGTGA	GAGTGA	-475
SLC2A1	GxGTGA	GAGTGA	-419
CITED2	GxGTGA	GAGTGA	-403
P4HA1	GxGTGA	GAGTGA	-99
ETS-1	GxGTGA	GAGTGA	-443
Alpha-1B Adrenergic Receptor	GCxTGA	GCCTGA	-271
	GCxTGA	GCCTGA	-89
SLC2A1	GCxTGA	GCCTGA	-231
Phosphofructokinase L	GCxTGA	GCCTGA	-380
GAPDH	GCxTGA	GCCTGA	-495
Enolase 1	GCxTGA	GCCTGA	-265
VEGF	GCGTGx	GCGTGC	-477
	GCGTGx	GCGTGC	-264
	GCGTGx	GCGTGC	-125
SLC2A1	GCGTGx	GCGTGC	-13
PFKFB3	GCGTGx	GCGTGC	-469
Phosphofructokinase L	GCGTGx	GCGTGC	-252
GAPDH	GCGTGx	GCGTGC	-394
	GCGTGx	GCGTGC	-314
	GCGTGx	GCGTGC	-180
Phosphoglycerate kinase 1	GCGTGx	GCGTGC	-235
IGFBP-1	GCGTGx	GCGTGC	-85
FLT-1	GCGxGA	GCGGGA	-339
	GCGxGA	GCGGGA	-23
Adrenomedullin	GCGxGA	GCGGGA	-500
Endothelin-1	GCGxGA	GCGGGA	-451
PFKFB3	GCGxGA	GCGGGA	-186
Phosphofructokinase L	GCGxGA	GCGGGA	-127
Aldolase A	GCGxGA	GCGGGA	-469
	GCGxGA	GCGGGA	-204

Gene	Motif	Consensus	Position
	GCGxGA	GCGGGA	-175
	GCGxGA	GCGGGA	-120
	GCGxGA	GCGGGA	-106
	GCGxGA	GCGGGA	-99
	GCGxGA	GCGGGA	-33
GAPDH	GCGxGA	GCGGGA	-51
Phosphoglycerate kinase 1	GCGxGA	GCGGGA	-77
P4HA1	GCGxGA	GCGGGA	-259
ETS-1	GCGxGA	GCGGGA	-107
	GCGxGA	GCGGGA	-76
	GCGxGA	GCGGGA	-68
Transferrin Receptor	ACGTGx	ACGTGr	-103
Alpha-1B Adrenergic Receptor	ACGTGx	ACGTGr	-84
PFKFB3	ACGTGx	ACGTGr	-107
Phosphofructokinase L	ACGTGx	ACGTGr	-287
Enolase 1	ACGTGx	ACGTGr	-69
Lactate Dehydrogenase A	ACGTGx	ACGTGr	-366
	ACGTGx	ACGTGr	-181
	ACGTGx	ACGTGr	-83
CITED2	ACGTGx	ACGTGr	-180
VEGF	ACGTxA	ACGTrA	-310
FLT-1	ACGTxA	ACGTrA	-112
Lactate Dehydrogenase A	ACGTxA	ACGTrA	-46
IGFBP-1	ACGTxA	ACGTrA	-264
Transferrin	ACGxGA	ACGrGA	-73
Ceruloplasmin	ACGxGA	ACGrGA	-165
PFKFB3	ACGxGA	ACGrGA	-258
IGFBP-1	ACGxGA	ACGrGA	-195
Transferrin Receptor	GCGTxA	GCGTnA	-366

Table 10. ScanPro analysis: "one-mistake-containing" exact results. Twenty-three genes were identified presenting one "mistake" inside one of the two expected HIF-1 binding site. The motif, as well as the derived consensus sequence, is presented as well as its position with respect to the Transcription Start Site.

Quite interestingly, by allowing the variability of one nucleotide inside the hexamer (an approximation that is present, often even at a higher range, in almost all the available predictive tools), the number of positive results found by ScanPro increased to 25, including 18 genes that had not been previously considered, the only gene not found to contain an HIF-1 binding site being the Intestinal Trefoil Factor.

The location of the "mistake" insertion is quite equally distributed among the six positions of the hexamer, although there seems to be a weak overabundance in the second and in the last nucleotide. It is useful to underline, however, that due to the fact that some inserted mistakes correspond to IUPAC characters used to define ambiguities (see the end of Table 10), the number of exact results would be even greater than the seven matches previously reported on the "forward" DNA strand and the five reported on the "reverse" strand.

The identified locations of HIF-1 binding sites are spread all over the (-500, -1) region of each gene, with respect to the Transcription Start Site as defined by EnsEMBL, although there seems to be some representing a preferential DNA-protein interacting point (Fig.24).

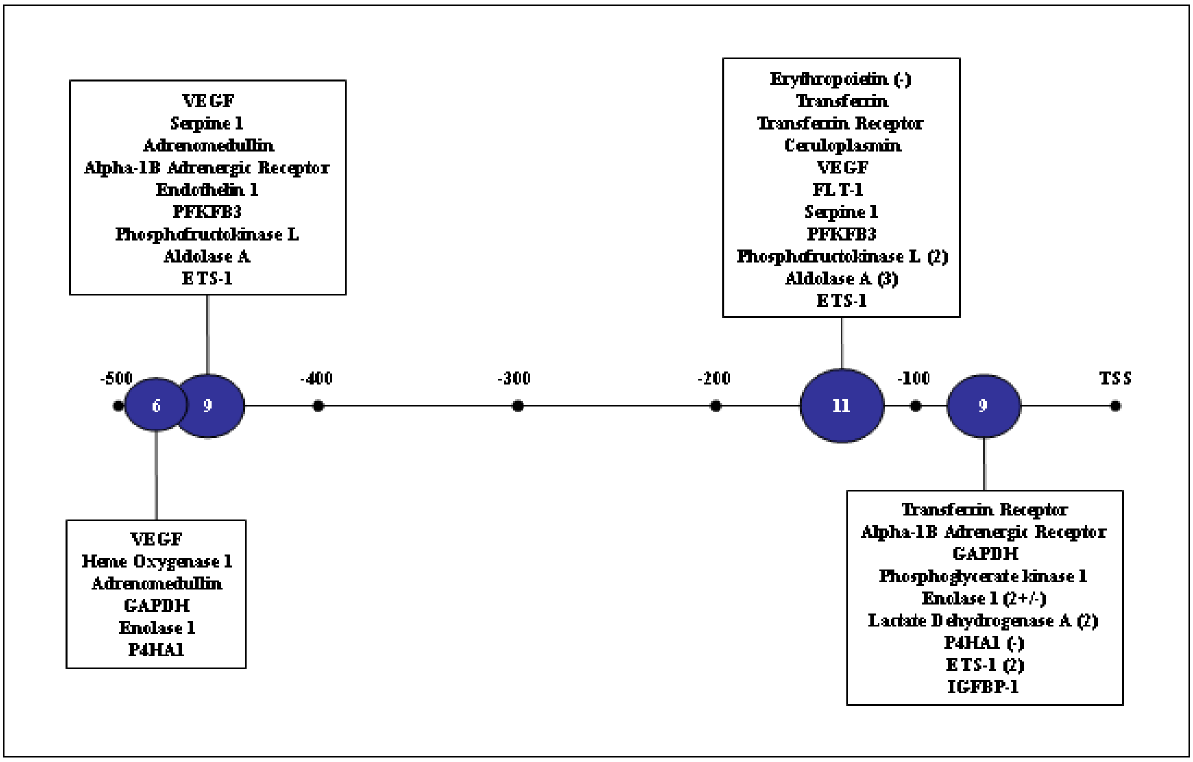


Figure 27. Location of the four more abundant TFBS identified among HIF-1 regulated genes.

Eleven different genes present a binding site in the (-140, -120) region, while nine other genes (only two present also in the previous location) bind DNA in the (-80, -70) region. It is then possible to define a first, huge preferential site for HIF-1 binding to its DNA targets in the (-140, -70) region, where 18 out of 26 regulated genes are able to undergo this interaction.

A second region that shows a relevant abundance of binding sites is the (-500, -460) region, with a peak of nine genes presenting the TFBS around the (-470, -460) position, and six other (only two present also in the previous location) in the (-500, -490) position. This second site, although slightly less involved in HIF-1 binding to DNA than the first one, still represents another preferential location for 13 out of 26 genes.

Globally, these two sites account for the interaction between the TF and 21 of its 26 target genes, with 10 of these presenting both the binding sites, while 8 exclusively depend on the first one and 3 on the second one.

Between these two regions that seem to play a major role in a large subset of genes, some other TFBS-rich locations can be found to be conserved in groups of five-seven genes. The common feature of all these sites, both the widely diffused and the more specific ones, is the spacing of ~40 bp between each element and its neighbours, almost suggesting some sort of steric hindrance and 3D requirement for the TF-TFBS interaction to occur. Finally, it is also possible to hypothesize different levels of involvement of HIF-1 binding sites depending on the kind of required regulation, with the two widely diffused binding sites being probably necessary for a more general, massive action of HIF-1, and the other that may be involved in more accurate and specific regulatory events.

Along with these motifs corresponding to the expected HIF-1 binding sites, ScanPro identified also other conserved motifs whose nature was evaluated as well.

I looked for conserved motifs with the same basic structure of the HIF-1 binding site, that is conserved motifs with size fixed to "6" and number of mutations to "1", in order to compare two similar sets of ScanPro results. To simplify the analysis and to reduce the number of available results, I also imposed the discovered conserved elements to be present in at least 22 of the 26 HIF-1 regulated genes.

Motif	**Binding Factor**	**Function**
CCCAGG	TOPORS (T04734)	tumor protein p53-binding protein/topoisomerase I binding; might be an important transcriptional regulator for lung cancer-associated genes including E-cadherin and talin
CCTGCC	AREB6 (T00625)	encodes a human zinc finger transcription factor that represses T-lymphocyte-specific IL2 gene (MIM 147680) expression by binding to a negative regulatory domain 100 nucleotides 5-prime of the IL2 transcription start site
CCTGGG		
CTGCCC	*AhR:arnt (T01795)*	*HIF-1 beta subunit*
GCAGAG	STAT5A (T04683)	member of the STAT family of transcription factors; this protein is activated by, and mediates the responses of many cell ligands, such as IL2, IL3, IL7 GM-CSF, erythropoietin, thrombopoietin, and different growth hormones. Activation of this protein in myeloma and lymphoma associated with a TEL/JAK2 gene fusion is independent of cell stimulus and has been shown to be essential for the tumorigenesis
GCTCCC	Unknown	Unknown; not a repeated or low complexity sequence

Motif	Binding Factor	Function
GCTGGG	PPARG (T03731)	member of the peroxisome proliferator-activated receptor (PPAR) subfamily of nuclear receptors. PPARs form heterodimers with retinoid X receptors (RXRs) and these heterodimers regulate transcription of various genes. PPAR-gamma has been implicated in the pathology of numerous diseases including obesity, diabetes, atherosclerosis and cancer. Multiple transcript variants that use alternate promoters and splicing have been identified for this gene.

Table 11. ScanPro analysis: other conserved elements. Seven motifs were found to be conserved in at least 22 genes regulated by HIF-1. For each motif are reported the known binding factor and the associated function (if any).

By looking at the results (Table 11), it is interesting first of all to observe that the fourth conserved motif is nothing else than another HIF-1 binding site recorded in Transfac, and in particular it corresponds to the hexamer that comes immediately after the 2 binding sites that were initially searched with ScanPro (thus explaining its absence from the first set of results).

The fifth, GCTCCC, could not be associated to any known Transfac TFBS; in the meantime, no match was found with any repeated sequence or low complexity element. It is not possible, yet, to affirm if this motif really corresponds to an unknown TFBS discovered by ScanPro or if its role goes beyond that function, but it is quite evident that its conservation in the regulatory regions of more than twenty genes probably underlies some biological function.

The five other results, finally, are very appealing for two main reasons. The first one is an algorithmical consideration: AREB6 and STAT5A binding sites were among the five best results that YMF produced when analyzing the HIF-1 dataset, thus confirming once more ScanPro predictive capability. The second one is an important biological consideration: all the TFs that bind these motifs are involved in immune response, in particular in Retinoic Acid and Interleukins mediated pathways (as was already reported in Table 5 at the end of the analysis of the function of HIF-1 target genes described by Transfac). Along this, there is also a strong involvement of these TFs in tumorigenic events, a fact that was recently demonstrated also for HIF-1 (Ohh M et al, Nat Cell Biol 2000; Cockman ME et al, J Biol Chem 2000; Krieg M et al, Oncogene 2000; Clifford SC et al, Hum Mol Genet 2001; Pennacchietti S et al, Cancer Cell 2003; Fels DR and Koumenis C, Trends Biochem Sci 2005). The exact molecular mechanisms used by hypoxia to unleash invasive and metastatic potential of tumour cells are largely unknown, but the accumulated evidence is starting do define a list of candidate genes regulated in a coherent way by a group of TFs whose interactions will have to be further investigated in the future, extending the search of this common set of TFBS to a broader list of genes.

DISCUSSION

A major effort in building the functional map of the human genome is its cross match with the transcriptional units. The 5' end boundary of each unit is defined by the promoter and the 3' end by the 3'UTR, with no compulsory need for coding potential. Considering the currently defined characters as contained in the EnsEMBL Genome Browser, it is critical to define other features and events such as alternative start sites and alternative splicing products, noncoding transcripts, TFBS, untranslated terminal repeats, and the role of other repeated sequences.

During my PhD I had the possibility to adress most of these topics.

Starting from the high-throughput production of human full-length cDNA libraries, I had the opportunity to evaluate and apply many tools to the analysis and functional annotation of cDNA sequences. The lack of publicly available tools allowing to perform some of the required bioinformatics analysis led me to interacting with informaticians, ad eventually to develop and implement an integrated computational system which is currently available for download and represents a useful and quite complete tool for the functional annotation of biological sequences.

The results of this first part of my PhD activity allowed me first of all to define the "LNCIB 5.8K Unique cDNAs collection" and, lately, to identify a subgroup of 342 clones corresponding to putative new genes, both coding and non-coding.

The availability of cDNA sequences, particularly from full-length cDNAs, certainly represents a very precious tool to improve the quality of genome annotation, as these sequences enable to define the precise protein coding parts of the genome and, in conjunction with the genomic counterpart, also to define the composition of exons in alternatively spliced transcripts of the same gene.

Both the sequence and the chromosomal location of genes constitute a very important piece of information supportive in the process of defining and analyzing candidate disease genes.

The everlasting, ongoing discovery of new genes or transcripts from the analysis of cDNA sequences also demonstrates the great usefulness of this approach in the identification of genes that would not otherwise be discovered in genomic sequences, thus indicating the need for caution when using *ab initio* predictions as the primary source for genome annotation.

All these considerations were clearly demonstrated during the analysis of the "LNCIB 5.8K Unique cDNAs collection". By applying a method allowing to enrich the full-length population of cDNA libraries (Carninci et al, Genomics 1996), followed by a simple additional procedure based on size selection of cDNAs >1500 bp applied to the library construction, it was possible to isolate from the three tissues and from the cell-line that were used for RNA extraction a total of ~5300 unique genes (~5800 considering the transcriptional spicing variants), representing almost one-third of the already known human cDNAs and one-fourth of the currently predicted human transcripts. More strikingly, the functional annotation of these cDNA sequences revealed that a subset of 342 poorly annotated sequences presented a correct gene structure, with a canonical exon–intron alternation, thus suggesting that this group could therefore represent candidates for new genes, confirming in the meantime what has already been reported about the limited predictive ability of the existing gene-prediction algorithms. Since then, new cDNA libraries were produced and 20,266 good quality sequences were analyzed with our integrated computational system and are now stored in TSDB, raising the number of clones pointing to unique Unigene clusters to 15,000, corresponding to 9404 unique LocusLink IDs. These results give a further demonstration of the usefulness of the developed analytical system; however, the latter sequences have not been fully studied, yet, and will have to be analyzed in a way similar to that applied for identifying the first 342 unknown clones.

A self-to-self hybridization cDNA microarray experiment confirmed that 260 out of 342 unknown genes were indeed expressed in the examined cell-lines or tissues used as source of RNA for the hybridization. The majority of these cDNAs derived from the human fetal brain or the human placenta libraries, thus confirming that the expressed cDNAs correspond to genes expressed during human development.

A second experiment of gene expression profiling was performed by using 30 human advanced ovarian cancer samples and four cell lines obtained by various treatments of the original IOSE-hTERT_INT cell line, and this time 289 unknown clones qualified as expressed.

By comparing the two lists of expressed genes, it was possible to identify a group of 236 unknown cDNAs that were expressed in both experiments, while 24 were expressed only in the self-to-self hybridization and 53 only in the gene expression profiling of the 30 human advanced ovarian cancers, thus representing a subgroup of interesting genes that could be further analyzed to study their possible involvement in ovarian cancer. Interestingly they could be clustered into different subgroups of co-expressed clones showing a common expression profile in the different samples. Further analysis of their genomic regions for the presence of TFBS and the identification of common motifs will possibly explain whether they are subjected only to co-expression or co-regulation. Construction of clusters of unknown genes and their comparison to known genes with similar expression profiles and shared TFBS will undoubtedly increase the knowledge of their functions.

Given the particular developmental stages of the tissues from which the RNA used for cDNA libraries preparation was extracted, it is quite possible that the 29 remaining transcripts whose expression could not be demonstrated correspond to tissue-specific, rare genes (or transcripts variants) that are very difficult to be observed during functional assays.

The 342 unknown genes were then tested for their coding potential and only 12 (of which 10 expressed in the previous cDNA microarrays experiments) showed a significant coding potential on the basis of the observation of the peculiar evolutionary dynamics of coding with respect to noncoding regions. As these clones fall in intergenic regions they reliably represent novel protein-coding genes whose general structure will be established by full-length clone sequencing. This hypothesis was also strongly supported by the comparison performed against the entries in the DBTSS database; in fact, only 1 of these 12 cDNAs had matches with human and mouse entries, and in particular matches were with hypothetical proteins, thus confirming that all these clones, including the latter, still have an unknown functional role in both species.

Finally, relevant information on the biological role of these unknown clones was given by transfections experiments, these data representing an initial survey for future functional studies. For six of the eight clones that were transfected in U2OS cells it was in fact possible to identify different interesting phenotypes, as followed by the change in the distribution of GFP used as co-transfection marker. These effects were quite dramatic in the case of clone 5000FJE08. It is not yet possible to understand if these effects are due to the translation of the clones or if they act in other ways, for example as antisense regulators. Bioinformatics analyses seem to demonstrate that the 12 unknown potentially coding clones indeed have coding capabilities, but further studies such as in vitro translations should be performed to confirm this hypothesis definitely. For this reason, I'm currently setting up the biological experiments that will be needed to test more accurately the functional role of the 12 novel protein-coding genes (in particular for the one leading to the most striking effect), while examining different possibilities (both available and still to develop) to start defining the putative biological role of the remaining novel non-coding human genes that I identified. It will be particularly interesting to focus the attention on transcripts derived from intronic regions of protein coding genes, as this class of transcripts seems to be quite abundant, and on the methods required for their isolation, amplification and cloning.

A recent work (Cheng et al, Science 2005) gives in fact a very interesting insight on these aspects, and introduces features that will certainly have a deep impact in the future of the human transcriptome analysis. It seems, from the analysis of the transcriptional maps of 10 human chromosomes, that only 19.4% of the transcribed regions of the human genome give rise to classical polyadenylated transcripts, with 36.9% (called *bimorphic*) existing also in the non-polyadenylated form and 43.7% (many with an exclusive nuclear location) existing ONLY in the latter status. The function of more than 50% of transcripts belonging to the first class is poorly characterized, while for the other two categories it is almost completely unknown. It seems that almost 50% of unannotated cytosolic poly A+ transcription is derived from the intronic regions within genes, as observed for some of the unknown member of LNCIB cDNA collection; despite the new origin of these transcripts, there is however strong evidence supporting the existence of canonical splicing events that lead to mature transcripts with coding potential for < 100 aa proteins. The presence of such a large portion of bimorphic transcribed sequences suggests that novel regulatory mechanisms may be involved in the identification of transcripts whose polyadenylation states are altered as means of regulation, as many of the detected bimorphic sequences are well-characterized coding genes found on the 10 analyzed chromosomes, and the evaluation of the protein coding potential of poly A- transcribed sequences is currently awaiting efficient methods to isolate, amplify and clone these types of transcripts.

In conclusion, the accumulated evidence on the characterization of these 342 clones firmly points to their definition as novel genes, only a minor fraction of them being protein coding. These results also suggest that these genes may be human-specific, thus confirming the usefulness of preparing cDNA libraries from less common tissues (like human fetal brain, used in this work) to identify genes that are expressed only in particular tissues during particular developmental stages. However, it is not yet possible to define their functional role, i.e., if they work with an antisense mechanism or by using other, less defined, noncoding methods, and future analysis will be needed.

The experimental confirmation that these unknown cDNAs could indeed represent novel genes was obtained by producing cDNA microarrays containing the whole unique cDNAs collection and testing the expression status of the respective transcripts in many cell-lines and human tissues, both physiological and pathological. The necessity to interpret gene expression results, and in particular to understand which of the generated clusters of co-expressed genes may indeed contain co-regulated genes, drove me to start taking into consideration the problem of TFBS identification, finally leading to the development of a new algorithm for TFBS search and to the birth of ScanPro.
In fact the further characterization of unknown cDNAs is being provided by the analysis of their genomic neighbourhood and by the identification of conserved elements involved in regulatory processes. The results presented in this thesis show how the promoter prediction programs that are currently being used on human sequences are indeed very useful while studying groups of promoters and regulatory regions associated to a certain amount of available information, while they tend to fail, or at least to perform not as well, in identifying unknown motifs. A significant improvement to this field has been provided by the application of phylogenetic footprinting, as the conservation of non-coding regions during evolution is supposed to be associated to functional conservation. However, as observed during the comparison of the human genome with other

vertebrate genomes (i.e. mouse and rat), the similarity is quite often limited to the coding regions and to small nearby non-coding regions. As suggested by Cawley et al (Cawley S et al, Cell 2004) these last regions should in fact be no longer considered as the promoters containing the exclusive, exhaustive binding sites responsible of the fine tuning of gene expression. This work, in fact, demonstrated that only ~22% of the TFBS for some of the most important TFs mapping to chromosomes 21 and 22 are located 5' to known genes, the great majority being instead distributed both in intronic locations and in intergenic regions. This is in line with the fact that current, traditional knowledge on transcription regulation is limited to a rather restricted number of TFs, that have been mapped to quite close genomic locations with respect to TSS and that, thus far, have not been able to give full satisfactory explanation to gene expression regulation questions. Certainly the network of TFs interactions is far from being clearly understood, and this is obviously one important source of ignorance. However it is also certain that the regions where binding sites have been searched thus far represent only a subset of those that should be investigated.

ScanPro, the algorithm for TFBS search presented in this thesis, tries to solve the problem by applying a *de novo* approach that performs a blind analysis of the examined sequences, preventing bias that may arise by focusing the attention on *a priori* knowledge. The validation of ScanPro was performed with a conservative approach, by considering the (-500, -1) regions that I have just admitted to be far from being the best target for this analysis, but this was a necessary compromise in order to compare it with other existing tools.

Given that premise, however, Scan Pro was able, along with identifying the expected TFBS in the HIF-1 dataset used for its validation, to also find other conserved elements that were shown to correspond to known TFBS. On the other hand the referenced algorithms, used in the comparison with ScanPro, could not find such TFBS.

More interestingly, the recurrence of these sites allowed me to define subgroups of genes belonging to common functional pathways. Understanding whether these subgroups are made of genes that are really co-regulated will require further studies like, for instance, cDNA microarrays experiments. In general, cDNA microarrays experiments performed at LNCIB constitute the embodied source of lists of genes to be examined using ScanPro to increase the degree of biological knowledge available during the interpretation of ScanPro output, a very important aspect required for distinguishing true positive results from computational noise.

Most importantly, along with algorithmical information my analysis also emphasized the role that HIF-1, along with other transcription factors, may play in pathological processes such as inflammatory response and tumorigenesis. The molecular mechanisms occurring during hypoxia are not well elucidated and it is certainly not possible, yet, to propose an exhaustive general model.

However, future work may find helpful to consider some of the ideas raised by this first analysis with ScanPro.

In order to make ScanPro a general tool, the next step will consist in extending the size and the complexity of the analyzed regions, the final goal being the application of ScanPro to entire gene sequences. This is indeed a difficult problem to solve, from the biological point of view. While working without any previous knowledge, either concerning the gene list composition or the length and structure of the expected TFBS, it is in fact necessary to fix a motif length and a number of mutations in order to allow ScanPro to generate some output. This situation will probably require the choice of at least 2-3 different settings for the algorithm, and the different results will need to be merged in order to gain significance. Along that, also the statistical validation of the obtained results is going to be one of the major improvements that I will need to introduce.

Meanwhile I have started to analyze a new group of genes, the p53 dataset. Along with the major importance that this pathway holds *per se*, there are two other reasons that led to its selection. First of all, p53 constitutes a central theme of research interest at LNCIB: along with great knowledge concerning its mechanism of action there is also the availability of many reagents that could be used to validate ScanPro predictions. The second reason, centred on available knowledge of the associated dataset, is its connection with the HIF-1 pathway.

The involvement of HIF-1 in tumorigenesis was firstly demonstrated some years ago (Ohh M et al, Nat Cell Biol 2000; Cockman ME et al, J Biol Chem 2000; Krieg M et al, Oncogene 2000; Clifford SC et al, Hum Mol Genet 2001) when it was shown that pVHL (von Hippel-Lindau tumour suppressor protein), that usually regulates physiological levels of HIF-1 by inducing its ubiquitinylation-mediated degradation, when inactivated led to constitutively high HIF levels and to the expression of many oxygen-regulated genes that in the end caused the formation of highly vascularized hemangioblastoma tumor, the most frequent manifestation of hereditary VHL disease, or of renal cell carcinoma. From a molecular point of view, the constitutive activation of anaerobic metabolism in tumor cells (Warburg effect) seems to provide an explanation for the cell autonomous response to oxygen deficiency, with HIF-1 function in hypoxic tumor adaptation that putatively involves VEGF-mediated angiogenesis and also increased glycolysis, pH buffering, and probably other key steps in tumor progression. Finally, the discovery of a direct interaction and regulatory effect of p53 on HIF-1 was recently reported (Fels DR and Koumenis C, Trends Biochem Sci 2005). These two TFs, that at first sight behave in an antithetical way, seem indeed to be involved in the same regulatory circuitry, as p53 was found to bind the ODD domain of HIF-1, thus inhibiting its transactivation and stimulating its degradation. This seems to occur under quite stringent hypoxic or anoxic conditions, when p53 accumulates and, after reaching a threshold level, directly interacts with HIF-1. Usually, under normoxic conditions (e.g. near a tumor blood vessel), levels of both p53 and HIF-1a are on the other hand low because of proteasome-mediated degradation. Further away from the vessel, owing to oxygen-diffusion limits, mild hypoxia develops that activates the HIF-1-dependent angiogenic program but is not stringent enough to induce p53 accumulation. Then, in the end, the threshold is attained and the interaction takes place.

The study of the p53 dataset is extremely interesting as, by now, there is no clear definition of the molecules that could co-participate in this very important tumour development control. It is also going to represent a further, significant step towards the complete validation of ScanPro as its significant complexity, both represented by the length of the palindromic motif and by the degree of variability in additional positions of the TFBS, represents a hard testing ground from the computational point of view. Finally, the availability of a tumor gene expression profiling DB at LNCIB will also allow me to verify the results that I'm starting to obtain from the analysis of a list of 82 p53-regulated genes (data not shown in this thesis), along with the relationship existing between p53 and HIF-1 in complex biological processes such as inflammatory response and tumorigenesis, as stated before.

The chromosome-wide or genome-wide approaches for TFBS search, performed without considering gene expression results but including phylogenetic footprinting data, and the motif search based on previous gene expression experiments, will probably be held in parallel, as I consider them to be complementary and to be necessary to draw the attention to different regulatory events.

In conclusion, I consider that the obtained results firmly suggest that ScanPro can be proposed as an accurate and efficient algorithm for the *de novo* discovery of human TFBS. In fact, in the HIF-1 dataset used for ScanPro validation, 25 out of 26 genes regulated by HIF-1 were identified, only one being missed, a result that could not be achieved with any of the other tested promoter prediction programs. The success rate of ScanPro is further reinforced by the fact that additional conserved elements, both known and unknown, were identified during this analysis, increasing the predictive capacity of ScanPro and allowing further comparison with other predictive programs.

Although considering that different improvements need to be introduced in the analysis, ScanPro has already laid the foundation as a successful tool in gene expression regulation.

BIBLIOGRAPHY

Allemann RK and Egli M. DNA recognition and bending. Chem Biol, 4 (9): 643-650 (1997).

Altschul SF et al. Basic local alignment search tool. J Mol Biol, 215 (3): 403-410 (1990).

Auffray C and Rougeon F. Purification of mouse immunoglobulin heavy-chain messenger RNAs from total myeloma tumor RNA. Eur J Biochem, 107 (2): 303-314 (1980).

Bajic VB et al. Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates. J Mol Graph Model, 21 (5): 323-332 (2003).

Bajic VB et al. Promoter prediction analysis on the whole human genome. Nat Biotechnol, 22 (11): 1467-1473 (2004).

Baross A et al. Systematic recovery and analysis of full-ORF human cDNA clones. Genome Res, 14 (10B): 2083-2092 (2004).

Basrai MA et al. Small open reading frames: beautiful needles in the haystack. Genome Res, 7 (8): 768-771 (1997).

Bird AP. CpG islands as gene markers in the vertebrate nucleus. Trends Genet, 3: 342-347 (1987).

Boffelli D et al. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science, 299 (5611): 1391-1394 (2003).

Boom R et al. Rapid and simple method for purification of nucleic acids. J Clin Microbiol, 28 (3): 495-503 (1990).

Burge C and Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol, 268 (1): 78-94 (1997).

Burke J et al. d2_cluster: a validated method for clustering EST and full-length cDNA sequences. Genome Res, 9 (11): 1135-142 (1999).

Carninci P et al. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. Genomics, 37 (3): 327-336 (1996).

Carninci P et al. High efficiency selection of full-length cDNA by improved biotinylated cap trapper. DNA Res, 4 (1): 61-66 (1997).

Carninci P et al. Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA. Proc Natl Acad Sci USA, 95 (2): 520-524 (1998).

Carninci P et al. The transcriptional landscape of the mammalian genome. Science, 309 (5740): 1559-1563 (2005).

Castrignano T et al. CSTminer: a web tool for the identification of coding and noncoding conserved sequence tags through cross-species genome comparison. Nucleic Acids Res, 32 (Web server issue): W624-627 (2004).

Cawley S et al. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell, 116 (4): 499-509 (2004).

Cheng J et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science, 308 (5725): 1149-1154 (2005).

Chiu C et al. Molecular evolution of the HoxA cluster in the three major gnathostome lineages. Proc Natl Acad Sci USA, 99 (8): 5492-5497 (2002).

Chomczynski P and Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem, 162 (1): 156-159 (1987).

Clifford SC et al. Contrasting effects on HIF-1α regulation by disease-causing pVHL mutations correlate with patterns of tumourigenesis in von Hippel-Lindau disease. Hum Mol Genet, 10 (10): 1029-1038 (2001).

Cockman ME et al. Hypoxia inducible factor-α binding and ubiquitylation by the von Hippel-Lindau tumor suppressor protein. J Biol Chem, 275 (33): 25733-25741 (2000).

Cooper GM et al. Characterization of evolutionary rates and constraints in three Mammalian genomes. Genome Res, 14 (4): 539-548 (2004).

Dalla E et al. Discovery of 342 putative new genes from the analysis of 5'-end-sequenced full-length-enriched cDNA human transcripts. Genomics, 85 (6): 739-751 (2005).

D'Alonzo RC et al. Physical interaction of the activator protein-1 factors c-Fos and c-Jun with Cbfa1 for collagenase-3 promoter activation. J Biol Chem, 277 (1): 816-822 (2002).

Davuluri RV at al. Computational identification of promoters and first exons in the human genome. Nat Genet, 29 (4): 412-417 (2001).

Del Sal G et al. The CTAB-DNA precipitation method: a common mini-scale preparation of template DNA from phagemids, phages or plasmids suitable for sequencing. Biotechniques, 7 (5): 514-520 (1989).

Deng JS et al. Internalization of anti-nucleolin antibody into viable HEp-2 cells. Mol Biol Rep, 23 (3-4): 191-195 (1996).

Dermitzakis ET et al. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature, 420 (6915): 578-582 (2002).

Diamond MI et al. Transcription factor interactions: selectors of positive or negative regulation from a single DNA element. Science, 249 (4974): 1266-1272 (1990).

Dickerson RE. DNA structure from A to Z. Meth Enzymol, 211: 67-111 (1992).

Down TA and Hubbard TJ. Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res, 12 (3): 458-461 (2002).

Dunham I et al. The DNA sequence of human chromosome 22. Nature, 402 (6761): 489-495 (1999).

Elliott DJ et al. An evolutionarily conserved germ cell-specific hnRNP is encoded by a retrotransposed gene. Hum Mol Genet, 9 (14): 2117-2124 (2000).

Elmasri R and Navathe SB. Fundamentals of Database Systems (4th edition). Addison-Wesley (2004).

Ewing B et al. Base-calling of automated sequencer traces using phred (part I&II). Genome Res, 8 (3): 175-185; 186-194 (1998).

Fels DR and Koumenis C. HIF-1α and p53: the ODD couple? Trends Biochem Sci, 30 (8): 426-429 (2005).

Ferg S. Modelling the time dimension in an entity-relationship diagram. Proceedings of the 4th International Conference on the Entity-Relationship Approach : 280-286. Spring S, Comp Soc Press (1985).

Fields C et al. How many genes in the human genome? Nature Genet, 7 (3): 345-346 (1994).

Fuchs T et al. The human olfactory subgenome: from sequence to structure to evolution. Hum Genet, 108 (1): 1-13 (2001).

Gerhard DS et al. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res, 14 (10B): 2121-2127 (2004).

Gibbs RA et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature, 428 (6982): 493-521 (2004).

Gregersen H and Jensen CS. Conceptual modelling of time-varying information. TimeCenter Technical Report TR-35, Department of Computer Science, Aalborg University (1998).

Griffiths-Jones S et al. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res, 33 (Database issue): D121-124 (2005).

Gumucio DL et al. Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human γ and ε globin genes. Mol Cell Biol, 12 (11): 4919-4929 (1992).

Gupta A. Life science research and data management – what can they give each other? SIGMOD RECORD, 33 (2): 12-14 (2004).

Hastings ML et al. Post-transcriptional regulation of thyroid hormone receptor expression by cis-acting sequences and a naturally occurring antisense RNA. J Biol Chem, 275 (15): 11507-11513 (2000).

Hattori M et al. The DNA sequence of human chromosome 21. Nature, 405 (6784): 311-319 (2000).

Hirotsune S et al. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. Nature, 423 (6935): 91-96 (2003).

Hla T and Maciag T. An abundant transcript induced in differentiating human endothelial cells encodes a polypeptide with structural similarities to G-protein-coupled receptors. J Biol Chem, 265 (16): 9308-9313 (1990).

Holmquist GP. Chromosome bands, their chromatin flavors, and their functional features. Am J Hum Genet, 51 (1): 17-37 (1992).

Hughes JD et al. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol, 296 (5): 1205-1214 (2000).

Inoue J et al. Multiple DNA elements for sterol regulatory element-binding protein and NF-Y are responsible for sterol-regulated transcription of the genes for human 3-hydroxy-3-methylglutaryl coenzyme A synthase and squalene synthase. J Biochem, 123 (6): 1191-1198 (1998).

Inman JT et al. A high throughput distributed DNA sequence analysis and database system. IBM System Journal, 40 (2): 464-488 (2001).

Kampa D et al. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res, 14 (3): 331-342 (2004).

Kapranov P et al. Large-scale transcriptional activity in chromosomes 21 and 22. Science, 296 (5569): 916-919 (2002).

Karp RM and Rabin MO. Efficient randomized pattern-matching algorithm. IBM J Res Dev, 31 (2): 249-260 (1987).

Katayama S et al. Antisense transcription in the mammalian transcriptome. Science, 309 (5740): 1564-1566 (2005).

Kawai J et al. Functional annotation of a full-length mouse cDNA collection. Nature, 409 (6821): 685-690 (2001).

Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res, 12 (4): 656-664 (2002).

Kiyosawa H et al. Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. Genome Research, 13 (6B): 1324-1334 (2003).

Kolbe D et al. Regulatory potential scores from genome-wide 3-way alignments of human, mouse and rat. Genome Res, 14 (4): 700-707 (2004).

Krieg M et al. Up-regulation of hypoxia-inducible factors HIF-1α and HIF-2α under normoxic conditions in renal carcinoma cells by von Hippel-Lindau tumor suppressor gene loss of function. Oncogene, 19 (48): 5435-5443 (2000).

Kulp D et al. A generalized hidden Markov model for the recognition of human genes in DNA. ISMB, 4: 134-142 (1996).

Kuwabara T et al. A small modulatory dsRNA specifies the fate of adult neural stem cells. Cell, 16 (6): 779-793 (2004).

Kwek KY et al. U1 snRNA associates with TFIIH and regulates transcriptional initiation. Nat Struct Biology, 9 (11): 800-805 (2000).

Lander ES et al. Initial sequencing and analysis of the human genome. Nature, 409 (6822): 860-921 (2001).

Lanz RB et al. A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. Cell, 97 (1): 17-27 (1999).

Lenhard B et al. Identification of conserved regulatory elements by comparative genome analysis. J Biol, 2 (2): 13 (2003).

Liang F et al. Gene index analysis of the human genome estimates approximately 120,000 genes. Nature Genet, 25 (2): 239-240 (2000).

Lipman DJ. Making (anti)sense of non-coding sequence conservation. Nucleic Acids Res, 25 (18): 3580-3583 (1997).

Lippman Z and Martienssen R. The role of RNA interference in heterochromatic silencing. Nature, 431 (7006): 364-370 (2004).

Liu et al. Conformational model for binding site recognition by the E.coli MetJ transcription factor. Bioinformatics, 17 (7): 622-633 (2001).

Loots GG et al. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science, 288 (5463): 136-140 (2000).

Loots GG et al. rVista for comparative sequence-based discovery of functional transcription factor binding sites. Genome Res, 12 (5): 832-839 (2002).

Lu XJ and Olson WK. Resolving the discrepancies among nucleic acid conformational analyses. J Mol Biol, 285 (4): 1563-1575 (1999).

Makeev VJ et al. Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. Nucleic Acids Res, 31 (20): 6016-6026 (2003).

Makeyev AV et al. A set of highly conserved RNA-binding proteins, alphaCP-1 and alphaCP-2, implicated in mRNA stabilization, are coexpressed from an intronless gene and its intron-containing paralog. J Biol Chem, 274 (35): 24849-24857 (1999).

Mattick JS and Gagen MJ. The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. Mol Biol Evol, 18 (9): 1611-1630 (2001).

Mattick JS. RNA regulation: a new genetics? Nat Rev Genet, 5 (4): 316-323 (2004).

Meister G and Tuschl T. Mechanisms of gene silencing by double-stranded RNA. Nature, 431 (7006): 343-349 (2004).

Mignone F et al. Computational identification of protein coding potential of conserved sequence tags through cross-species evolutionary analysis. Nucleic Acids Res, 31 (15): 4639-4645 (2003).

Montgomery MK. RNA interference: historical overview and significance. Methods Mol Biol, 265: 3-21 (2004).

Ohh M et al. Ubiquitination of hypoxia-inducible factor requires direct binding to the β-domain of the von Hippel-Lindau protein. Nat Cell Biol, 2 (7): 423-427 (2000).

Ohler U et al. Computational analysis of core promoters in the Drosophila genome. Genome Biol, 3 (12): RESEARCH0087 (2002).

Ohno S. Dispensable genes. Trends Genet, 1: 160-164 (1985).

Okazaki Y et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature, 420 (6915): 563-573 (2002).

Ota T et al. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet, 36 (1): 40-45 (2004).

Pavesi G et al. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res, 32 (Web server issue): W199-203 (2004).

Pevzner PA and Sze SH. Combinatorial approaches to finding subtle signals in DNA sequences. Proc Int Conf Intell Syst Mol Biol, 8: 269-278 (2000).

Ponger L and Mouchiroud D. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. Bioinformatics, 18 (4): 631-633 (2002).

Prasanth KV et al. Regulating gene expression through RNA nuclear retention. Cell, 123 (2): 249-263 (2005).

Pruitt KD et al. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res, 29 (1): 137-140 (2001).

Roest Crollius H et al. Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. Nature Genet, 25 (2): 235-238 (2000).

Rossant J and McKerlie C. Mouse-based phenogenomics for modelling human disease. Trends Mol Med, 7 (11): 502-507 (2001).

Salamov AA et al. Assessing protein coding region integrity in cDNA sequencing projects. Bioinformatics, 14 (4): 384-390 (1998).

Sanger F et al. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA, 74 (12): 5463-5467 (1977).

Sarafova S and Siu G. Precise arrangement of factor-binding sites is required for murine CD4 promoter function. Nucleic Acids Res, 28 (14): 2664-2671 (2000).

Scully KM et al. Allosteric effects of Pit-1 DNA sites on long-term repression in cell type specification. Science, 290 (5494): 1127-1131 (2000).

Segal E et al. From signatures to models: understanding cancer using microarrays. Nat Genet, 37 (Suppl): S38-45 (2005).

Semenza GL and Wang GL. A nuclear factor induced by hypoxia via de novo protein synthesis binds to the human erythropoietin gene enhancer at a site required for transcriptional activation. Mol Cell Biol, 12 (12): 5447-5454 (2002).

Shiraki T et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci USA, 100 (26): 15776-15781 (2003).

Sinha S and Tompa M. YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Res, 31 (13): 3586-3588 (2003).

Sleutels F et al. The non-coding Air RNA is required for silencing autosomal imprinted genes. Nature 415 (6873): 810-813 (2002).

Solovyev V and Salamov A. The Gene-Finder computer tools for analysis of human and model organisms genome sequences. ISMB, 5: 294-302 (1997).

Spellman PT and Rubin GM. Evidence for large domains of similarly expressed genes in the Drosophila genome. J Biol, 1 (1): 5 (2002).

Sterner DA et al. Architectural limits on split genes. Proc Natl Acad Sci USA, 93 (26): 15081-15085 (1996).

Storz G et al. An abundance of RNA regulators. Annu Rev Biochem, 74: 199-217 (2005).

Suzuki Y et al. DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. Nucleic Acids Res, 32 (Database issue): D78-81 (2004).

Tagle DA et al. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol, 203 (2): 439-455 (1988).

Taudien S et al. RUMMAGE – a high throughput sequence annotation system. Trends Genet, 16 (11): 519-520 (2000).

Terai G and Takagi T. Predicting rules on organization of cis-regulatory elements, taking the order of elements into account. Bioinformatics, 20 (7): 1119-128 (2004).

Tompa M et al. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol, 23 (1): 137-14 (2005).

Venter CJ et al. The Sequence of the human genome. Science, 291: 1304-1351 (2001).

Wasserman WW et al. Human-mouse genome comparisons to locate regulatory sites. Nature Genet, 26 (2): 225-228 (2000).

Waterston RH et al. Initial sequencing and comparative analysis of the mouse genome. Nature, 420 (6915): 520-562 (2002).

Wenger RH. Cellular adaptation to hypoxia: O_2-sensing protein hydroxylases, hypoxia-inducible transcription factors, and O_2-regulated gene expression. FASEB J, 16 (10): 1151-1162 (2002).

Wheelan SJ et al. Spidey: a tool for mRNA-to-genomic alignments. Genome Res, 11 (11): 1952-1957 (2001).

Wingender E et al. The TRANSFAC system on gene expression regulation. Nucleic Acids Res, 29 (1): 281-283 (2001).

Workman CT and Stormo GD. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. Pac Symp Biocomput, 467-478 (2000).

Xiang CC et al. Amine-modified random primers to label probes for DNA microarrays. Nat Biotechnol, 20 (7): 738-742 (2002).

Xie X et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature, 434 (7031): 338-345 (2005).

Xuan Z et al. Genome-wide promoter extraction and analysis in human, mouse, and rat. Genome Biol, 6 (8): R72 (2005).

Zalfa F et al. The fragile X syndrome protein FMRP associates with BC1 RNA and regulates the translation of specific mRNAs at synapses. Cell, 112 (3): 317-327 (2003).

Zhu Z et al. Discovering functional transcription-factor combinations in the human cell-cycle. Genome Res, 15 (6): 848-855 (2005).

ACKNOWLEDGMENTS

I am very honoured to thank Prof. Michael Q Zhang for accepting to be my external supervisor, his work has been a very important guide during my PhD experience.

I would like to thank Prof. Claudio Schneider for allowing me to work in his laboratory all over these years and for the precious feedbacks he gave me during all our discussions.

I would like to thank Dr. Marco Zantoni for all the work done with ScanPro and for being an exceptional colleague and friend. It has been a real pleasure to work together.

I would like to thank Dr. Michele Braidotti and Dr. Silvano Piazza for sharing with me everyday's work, for the help they provided during all these years and for all the coffees we had together. You are great!

I would like to thank all my colleagues and friends at LNCIB and SISSA/ISAS for their kindness.

I would like to thank Dr. James Reid for introducing me to bioinformatics … it's all your fault!

Finally, I would like to thank SISSA/ISAS for the scientific and economic support I received during my PhD years.