

Prediction, Annotation and Analysis of Human Promoters

Michael Q. Zhang

Cold Spring Harbor Laboratory

1 Bungtown Road, Cold Spring Harbor, NY 11724

INTRODUCTION

Since the celebrated discovery of Watson-Crick double-helix structure of DNA, it has taken 50 years for human genome to be sequenced. It may very well take another 50 years for the functional information to be fully decoded. Up till recently, genome research has mainly been focusing on coding regions, where the immediate questions are “where are the protein coding regions?” and “what are the functions of the gene products”. Increasingly, the field is advancing towards non-coding regions, where the central questions become “where are the regulatory regions?” and “how do they control gene expressions”. In 1961, Jacob and Monod published “On the regulation of gene activity” at the 26th Cold Spring Harbor Symposium on Quantitative Biology, in which some of the fundamental concepts of gene regulation were first elegantly formulated. Regulatory regions are most fundamental, because all the gene structures are defined by and recognized through the *cis*-elements in such regions; further more, what a gene does in vivo is intimately related to when, where and how much it is expressed. A phenotype, upon which the selection force is acting, is the integrated result of gene function and regulation. It is argued that the animal diversity is mainly due to the evolutionary expansion in regulatory complexity (Levine & Tjian 2003). Most regulations occur at the transcriptional level and the initiation of transcription is largely determined by the promoter located at the beginning of each gene, identification of promoters and *cis*-regulatory elements within them has become the prerequisite for understanding of gene regulation. For a few model organisms with compact genome (such as phage, bacteria and yeast), many of the gene regulatory pathways or networks have been worked out. But for mammalian systems, such as human, systematically identification of regulatory

regions and gene networks have turned out to be extremely difficult, largely due to the size and complexity of the genomes (hence, as a result, the diversity of the cell/tissue types and the complication of developmental stages).

Here I will outline our approaches to this problem. As genome research is data and technology driven, many approaches in the field can soon become obsolete once new or more data or technologies become available. I will try to state generic ideas and methodologies that may be evolving with or refined by new data or technologies. I will also try to point out open problems and to suggest new experiments to attack them.

***In silico* prediction of mammalian promoters**

Transcription of a eukaryotic protein-coding gene is preceded by multiple events; these include decondensation of the locus, nucleosome remodeling, histone modifications, binding of transcriptional activators (or derepressors) and coactivators to enhancers and promoters, and recruitment of the basal transcription machinery to form the preinitiation complex (PIC) at the core promoter. A core promoter is defined approximately as the DNA region (−40,+40) with respect to the transcriptional start site (TSS). It may contain the TFIIB recognition element (BRE) and the TATA-box at the 5'-end, the initiator (Inr) around the TSS and the downstream promoter element (DPE) at the 3'-end (see, *e.g.* Smale and Kadonaga 2003). Although, in a mammalian genome, distal enhancers/silencers can be 10 ~ 100 kb away from the target gene; most of the *cis*-regulatory elements are contained in a proximal promoter region of 0.5 ~ 2 kb in size. Putatively mapping of known transcription factor binding site (TFBS) density profile was originally used to develop the first computational promoter prediction program called *Promoterscan* (Prestridge 1995, see Fickett and Hatzigeorgiou 1997 for survey and

evaluation of earlier promoter prediction programs), later discriminative oligo-nucleotide based algorithms, such as *PromoterInspector* (Scherf *et al.* 2000), showed much improved performance.

We hypothesized that the molecular pattern recognition may be achieved by different molecular machinery with different resolutions at different scales (Zhang 1998a). An analogy would be that, if one tries to locate a landmark on earth from an airplane, one could use a coarse-grained tool to locate a regional landscape before zooming in with a finer mapping tool. Ideally, a coarse-grained promoter finder should be able to detect a chromatin and/or epigenetic landscape at the proximal promoter level (resolution < 2 kb). It could be an easier problem if one had 3D structural images (and this could happen within the next 10 years). With only the primary DNA sequences, one would have to use large-scale statistical features of those length characteristics. Fortunately, for human (or vertebrate), CpG islands can provide one such discriminative feature for at least 50% genes (Antequera and Bird 1993)! The human genome contains ~ 50,000 CpG islands, ~30,000 after repeatmasking and majority of these are near promoters. Computationally, a CpG island is defined (Gardinger-Garden and Frommer 1987) by a DNA region > 200 bp that has > 50% GC-content and > 0.6 ratio of CpG over expected CpG. Using this criteria, one would find ~345,000 CpG islands in the human genome. By detecting promoter associated CpG islands, we have developed an algorithm (called *CpG_Promoter*) for coarse-grained promoter mapping (Ioshikhes and Zhang 2000). Promoter associated CpG islands tend to be larger (0.5 ~ 2 kb), higher GC-content and the CpG ratio, other CpG islands are mostly associated with Alu repeats. Takai and Jones (2002) proposed a new definition: size > 500 bp that has > 55% GC-content and >

0.65 CpG ratio. Using this new criteria, one would find ~37,000 CpG islands. Later other CpG island based promoter prediction algorithms, such as *CpG+* (Hannenhalli and Levy 2001) and *CpGProD* (Ponger and Mouchiroud 2002), have also become available. We would like to see more large-scale experimental data, such as chromosome bandings, methylation patterns, histon modification profiles, Dnase hypersensitive sites, ChIP profiles and genomewide transcription reporter constructs. Integrating these data will allow better promoter landscape mapping algorithms to be developed.

For a finer promoter mapping, aiming at predicting the TSS with resolution < 100 bp, we developed an algorithm, called *CorePromoter* (Zhang 1998b), based on quadratic discrimination analysis (QDA) using position-dependent oligo-nucleotide features (these positions are designed to capture the known core-promoter elements). By combining a coarse-grained and a fine prediction tools, I demonstrated how the TSS could be precisely located for App gene (a 300 kb gene in chromosome 21) encoding Amyloid precursor (Zhang 2000). Instead of oligomers, *Eponine* uses a set of weight matrices in a hybrid machine-learning approach (Down and Hubbard 2002) to identify TSS. *Dragon Promoter Finder (DPF)* is an Artificial Neural Network (ANN) based algorithm which uses multiple sensors (promoters, exons and introns) to predict TSS (Bajic *et al* 2002).

As gene structures are often correlated (*i.e.* neighboring introns or exons can help predicting promoters as demonstrated in *DPF* above. See review by Zhang 2002a). We have developed *FirstEF* that integrates promoter, 5'UTR and first-intron information for predicting human first exons and promoters simultaneously (Davuluri *et al* 2001). There is an increasing evidence that transcription and splicing are coupled, we expect that promoter may influence the first donor site selection. Recently, *DPF* output and CpG

islands were integrated into a larger ANN program called *Gene Start Finder (DGSF)* to achieve a comparable promoter prediction in a test using chromosome 4, 21 and 22 (Bajic and Seah 2003). Although modern gene prediction programs, such as *Genscan* (Burge and Karlin 1997), try to predict first coding exons; *FirstEF* is the only program that is capable of predicting non-coding (untranslated) first exons. The most important open problems in promoter prediction are (1) how to improve accuracy on predicting promoters (or first exons) that are not CpG island associated; (2) how to predict alternative promoters and to predict multiple TSSs (a single promoter can regulate multiple start sites especially when the multiple start site downstream element (MED-1) is present (Ince and Scotto 1990).

Automatic construction of CSHL Mammalian Promoter Database reference system

As microarray expression data become prevalent, biologists often need to extract various sets of promoter sequences from clustered genes (Zhang 1999a). Originally, we developed *PEG* (Promoter Extraction from GenBank) using a set of accession numbers or ESTs as the input to facilitate the extraction of large sets of promoters (Zhang and Zhang 2001). When the nearly finished genome became available in April 2003 (Human built 33), we developed our automated annotation pipeline (an expert system) called *FexAnnotator* (First exon annotator, Davuluri *et al.* 2003), which can reduce false-positives and false-negatives from the *FirstEF* predictions by using existing knowledge in the public sequence database annotations (mRNA/EST and ENSEMBL genes). In this first pass annotation, we have ~53,000 first exons (including ~8,000 alternative first exons only annotated for the Refseq genes). The accuracy check shows that among ~10,000 experimentally verified first exons (such as those in *EPD* and in *DBTSS*), ~80%

were found within 500 bp of our pipeline predictions. Another check using known TFBSs in *TRANSFAC*, the density if these TFBSs are indeed concentrating within the vicinity of annotated core promoters.

For genome scale regulation studies, building a high quality promoter database, which allows easy and flexible data query or retrieval as well as on-the-fly analysis, is essential. Our *Saccharomyces cerevisiae* Promoter Database (*SCPD*, Zhu and Zhang 1999) has proved to be very instrumental for the yeast community. In order to better annotate the human promoters, we are currently building the *CSHL Mammalian Promoter Database*, which initially includes Homo sapiens Promoter Database (*HsPD*, based on Human built 33, April 2003), Mus musculus Promoter Database (*MmPD*, based on Mouse release Feb. 2003) and Rattus norvegicus Promoter Database (*RnPD*, based on Rat release Jan. 2003). A new pipeline has been developed (Z.Y. Xuan *et al.* unpubl.), in addition to ENSEMBL (Hubbard *et al.* 2002), it also makes use of results from *GenomeScan* (Yeh *et al.* 2001), *Fgeneh+* (Solovyev 2002) and *TwinScan* (Korf *et al.* 2001) in order to annotate promoters for potential novel genes (for further experimental validations). The new pipeline, taking advantage of cross-species comparisons, can automatically annotate multiple genomes in parallel on a Linux cluster (Figure. 1) and have been used to create the initial reference system for the *CSHL Mammalian Promoter Database* (Z.Y. Xuan, F. Zhao, *et al.* unpubl.). In this database, orthologous promoters will be linked so that a user can input a list of UnigeneIDs or Accession Numbers (from a clustered microarray data, say), specify the range of promoter region, extract orthologous promoter sequences, do motif finding on-the-fly; or select a gene of interest, do orthologous promoter alignment on-the-fly and look for conserved motifs (Figure 2).

Maintaining computability in addition to manual browsability will serve well to both computational and experimental biologists.

Functional curation of cell cycle transcription factors and their target genes

A promoter reference system created by automatic pipeline can insure completeness, it is consistent with most of the known information and also has reasonable accuracy. It must contain rich functional information (TFs, TFBSs, TSS, CpG islands) and links to other related databases and literature reference in order to be useful. Therefore, we are adding on top of *HsPD/MmPD/RnPD* with *TRED* (Transcription Regulatory Element Database, F. Zhao *et al.* unpubl.) (Figure 1), which allows semi-automated or even hand-curated information to be entered. Three most important issues every useful database must face to are (1) assign quality value to the raw record; (2) insure accuracy and usefulness; (3) open data disseminations. For (1), we have assigned different quality values to promoters and TFBSs according to how they were derived. For (3), we are discussing with NCBI (D. Lipman, pers. comm.) and EBI (E. Birney, pers. Comm.) on ways to incorporate our results into public databases. The most difficult and time-consuming task is (2), which involves hand-curation and outreach to transcription expert labs. We are initially focusing on cell cycle and cancer related TFs including their target genes, and will give authorship to related transcription labs that contribute data or expertise. Currently, out of 60,519 promoters (40,658 genes) in human part of *TRED*, only 2003 promoters (1853 genes) are in the best quality class (known and curated class). Other classes are: known but not curated, predicted based on Refseq, predicted based on other mRNAs, predicted based on other ESTs and purely predicted. As an example, for human E2F targets, *TRED* contains 233 promoters (182 genes) in the best quality class.

High throughput experimental validations

All computational predictions must be subject to experimental verifications and both positive as well as negative results are crucial feedbacks for further database and algorithm improvement. Lacking high throughput experimental validation has become the bottleneck in this feedback loop. As cDNA libraries become more saturating, novel gene finding has gradually shifted its paradigm from EST sequencing to computational prediction plus experimental validation (Das *et al* 2001, Guigo *et al* 2003). To validate first exons and TSSs, getting 5'-complete cDNAs are essential (Suzuki *et al.* 2000, Davuluri *et al* 2000). Recently using reporter construct, 5'-quality of Refseq and MGC clones have been randomly assayed for transcriptional activity of the upstream sequences (Trinklein *et al.* 2003).

In collaboration with McCombie and Hannon labs at CSHL on developing high throughput experimental genome annotation technologies, we have performed systematic 5'-RACE-PCR validation of 300 first exon predictions in 15 mouse tissue libraries (Baliya *et al.* 2003). We have selected the predicted exons in 5 categories according to having support evidence from: (a) EPD (this serves as the positive control), (b) Refseq, (c) more than two ESTs, (d) only one EST, (e) pure prediction. The success rates are 12/13, 17/27, 18/23, 28/169 and 16/68, respectively. Here ~ 25% predicted novel genes are likely to be real.

Working with Wang lab at U. of Chicago on developing GLGI (Generation of Longer 3'cDNA from SAGE Tag for Gene Identification, see Chen *et al.* 2003) -based genome annotation technology, we also obtained 57 positives from a test of 104 first

exon predictions in human tissues and 15 full length cDNAs were sequenced from 47 novel exon/SAGE-tag clones (S.M. Wang, pers. comm.).

To test promoter activities, we have collaborated with Stubbs lab at LLNL in annotating predicted genes/promoters in 800 kb region (containing 48 genes) of human ch19q13 using luciferase report system in addition to RT-PCR. Out of 38 tested predictions, 26 were tested positive (L. Stubbs, pers. comm. and see Figure 3).

These experimental exercises have demonstrated the validity of the large-scale computational prediction plus experimental verification approach for accurate genome annotation. It is also alarming that many previous false positives can be turned into true positives after more issues are tested or more sensitive experimental techniques are used (kapranov *et al.* 2003). The new challenge for computational biologists is how to recover false negatives; while for experimental biologists is how to prove a false positive!

Computational challenges in identification of *cis*-regulatory elements and transcriptional networks

Although most TFBSs are in the promoter region, many may be in the first intron (which can also be located by *FirstEF* prediction) and some may be in the 3'-flanking region (which can be located by EST/poly(A) mapping). There are also many distal enhancers/silencers/boundary elements that are so far away from the target genes, they are the most difficult to find. And even if they are found, linking to the correct target genes is still no easy task. We are focusing on proximal promoter region for *cis*-regulatory element discovery; many of the methods may also be applied to other regulatory regions once they are approximately localized (for example, by comparative

genomic analysis or by DNase hypersensitivity mapping or enhancer trapping technologies).

A. Computation-then- validation paradigm

Traditionally, identification of a *cis*-regulatory element is very laborious: collect known binding sites, build consensus or weight matrix and search for new loci. One cannot discover novel sites in this way. To study human cell cycle regulation, we have developed E2F *SiteScan* based on genetic algorithm trained on known sites in *TRANSFAC* and scanned ~5,000 promoters in the public database to identify more than 300 E2F targets, many of which were also validated by ChIP-PCR method (Kel *et al.* 2001). Since E2F motif was built mainly from known cell cycle genes, they may be biased as E2F also plays important roles in other biological pathways (such as apoptosis, DNA repair, *etc.*). By analyzing promoters from ChIP-PCR top candidates, we were able to identify novel E2F targets that do not have the conventional binding motif (Weinmann *et al.* 2001). But the scope with PCR is still very limited. When large-scale genome-wide data and technologies become available, one now is able to study TFBS in the whole genome together with their transcriptional readouts. It is expected that computational approaches are becoming more indispensable and will play more important roles in the future to come (Zhang *et al.* 2002, Zhang 2002b).

B. Large scale gene expression analysis

DNA microarray gene expression has become the widely used methods for studying gene regulation. It provides the direct readout of the cellular transcriptional programs. Interpretation of gene expression patterns by *cis*-elements and *trans*- factors, or conversely reconstruction of regulatory circuits from transcriptional responses is the main

challenge in the 21th century (Zhang 1999a, Banerjee and Zhang 2002). Using cluster analysis followed by motif searching of promoters of co-regulated genes, we were quite successful in identification of *cis*-elements involved in yeast cell cycle regulations (Spellman et al 1998, Zhang 1999b). By combining functional information, such as *MIPS* (Zhu and Zhang 2000) or *GO* (Chen *et al.* 2003), one can further select gene clusters that are not only co-expressed but also share significant number of genes involved in similar functional pathways or structural complexes.

Human *cis*-element detection is much more difficult due to much smaller signal-to-noise ratio (promoter region is much larger and uncertain, motifs are more degenerate, there are many repeats, *etc.*). Most commonly used motif finders, such as *Consensus* (Hertz *et al.* 1990), *MEME* (Beiley and Elkan 1994) and *Gibbs sampler* (Lawrence *et al.* 1993, Neuwald *et al.* 1995), assume a specific background model (*e.g.* Markov of order *k*). In order to increase specificity, we have developed a novel motif finding software package called *BEAST* (Binding Element AnalySis Tools, Hata and Zhang, 2003) that allows arbitrary background sequences to be the control set. The algorithm is based on exhaustive word counting strategy (allowing gap and reverse-complement, overlapping word is treated similarly as in van Helden *et al.* 2000). For each motif, the Fisher exact test (or chi-square test with Yates's correction) is used to evaluate *p*-value (with multiplicity correction) for the significance of motif association to the target (promoter) sequences against the background control. *BEAST* has been applied to microarray expression data from transcription factor knockout experiments (Chen *et al.* 2003), using the up regulated promoters, the down regulated or the combination as the target and using

the unchanged as the control. Combined with *GO* annotation (Ashburner *et al.* 2000), results agree well with the corresponding ChIP-chip analysis (data not shown).

BEAST was tested in detecting liver-specific promoter elements when a set of 35 proximal promoters of known liver specific genes was used for the targets and the pool of 1800 EPD promoters was used as the control. The HNF-1 motif YAMT..TTRA ($p=6.1 \times 10^{-12}$) was clearly identified on top of other putative motifs (Hata and Zhang 2003). The new challenge is to apply *BEAST* systematically to mammalian tissue expression data, using tissue-specific gene promoters as the targets and using the pool as the control, for discovering tissue-specific promoter elements. Future adaptation of *BEAST* with weight matrices should further improve its sensitivity for degenerate motifs or motif combinations.

C. Large scale chromatin localization analysis

Unlike the indirect co-regulation strategy above, ChIP-chip assay allows to detect TF binding targets in the whole genome by cross-linking protein to chromatin DNA *in vivo*. The first two human ChIP-chip experiments were done using a CpG island DNA chip (Weinmann *et al.* 2002) or using a Refseq genes promoter chip (Ren *et al.* (2002) to map E2F4 target genes.

In collaboration with Ren lab, we have used ChIP-chip assay to discover a global transcriptional regulatory role for c-myc in Burkitt's lymphoma cells (Li *et al.* 2003). We find that c-myc together with its heterodimeric partner, Max, occupy more than 15% of the gene promoters tested and they colocalize with TFIID in these cells, indicating a general role for over-expressed c-myc in global gene regulation of some cancer cells. One surprise from the promoter analysis is that many of the targets do not have the

conventional E-box, instead we find a novel motif **CGGAAG** by BEAST which is the most significant *cis*-element shared by large number of c-myc/Max binding target promoters (Hata *et al.* unpubl.). Furthermore, most of the elements are located near TSS (within 100 bp) and their positions are conserved among human, mouse and rat (data not shown). We are currently seeking experimental test for its functional relevance.

Recently there are two other motif detection algorithms suitable for ChIP-chip and expression data analysis. One is a word-based linear regression algorithm called *REDUCE* (Bussemaker *et al.* 2001) and another is a hybrid (word enumeration and weight matrix) greedy search algorithm called *MDscan* (Liu *et al.* 2002). Comparing to these, BEAST conveniently provides motif p-values and is more discriminative against a given background control set.

D. Comparative genomic analysis

Increasingly, comparative genomics has become very powerful method for detecting functional elements in non-coding regions. We began with a compative DNA sequence analysis of mouse and human protocadherin gene clusters in collaboration with experimentalists. The genomic organization of the human protocadherin α , β and γ gene clusters (designated *Pcdh α* , *Pcdh β* and *Pcdh γ*) is remarkably similar to that of immunoglobulin and T-cell receptor genes. The extracellular and transmembrane domains of each protocadherin protein are encoded by an unusually large “variable” region exon, while the intracellular domains are encoded by three small “constant” region exons located downstream from a tandem array of variable region exons. By comparing human draft and mouse BAC sequences, we were able to identify an alternative CpG island associated promoter in front of each variable exon in the α and γ gene clusters as

well as a highly conserved *cis*-regulatory element within the promoter (Wu *et al.* 2001). Later, it was further confirmed that these *cis*-elements are functionally important (Wang *et al.* 2002) and alternative promoter choice determines first intron splice site selection (Tasic *et al.* 2002).

To build our comparative genomics infrastructure, we carried out whole genome comparison between human and both (Celera and public) versions of mouse assemblies and published our *CSEdb* (*Conserved Sequence Element*) (Xuan *et al.* 2002). CSEs cover ~ 3% of the human genome. One third of these CSEs are related to known genes, some are related to other functional elements (such as RNA genes, antisense genes, *etc.*); but more than half are still functional unknown. Unknown CSEs provide excellent candidates for discovering novel genes or *cis*-regulatory regions. CSEs also allow us to arrive at another independent estimate of the number of human genes (~40,000).

Although comparative genomics has proved to be promising for discovering *cis*-regulatory regions (Pennacchio and Rubin 2001), because different promoter evolves with different rate, multiple species would have to be needed for narrowing down to short TFBSs. Initial success in yeast (Kellis *et al.* 2003, Cliften *et al.* 2003) may not directly translate in human, novel integrated approaches would have to be required to teeth out functional *cis*-elements even if the number of mammalian genomes were doubled.

E. Integration, combinatorial analysis and network reconstruction

Genomic data is noisy; the best weapon for combating noise is signal correlation analysis. Combinatorial interaction among TFs introduces correlation among their binding sites. Recently, there have been new motif finding algorithms, such as *CO-Bind* (GuhaThakurta and Stormo 2001), that are designed specifically for detecting correlated

motifs. Integration of evolutionary conservation with word-pair analysis can yield a better regression to expression data (Chiang *et al.* 2003).

Integrating ChIP-chip and expression data at the single motif level has recently attempted (Conlon *et al.* 2003). We have developed two methods for studying cooperativity by integrating ChIP-chip data and microarray expression data. For a given pair of TFs, A and B, the first method compares expression patterns of the targets of both TFs to that of A or B alone. If the former is more coherent (correlated), it is more likely that the two TFs are interacting in the transcription regulation of their common targets (Banerjee and Zhang 2003). The second method further integrates with promoter sequence analysis in order not only to infer the interacting TFs, but also to assign their corresponding binding sites by iteratively and exhaustively searching for significant TFs combinations and motifs combinations up to the triplet level (Kato *et al.* 2003). After analyzing over hundred TF ChIP-chip data (Lee *et al.* 2002), we were able to reconstruct the yeast cell cycle transcriptional regulation network so that (1) it extends the previous chain of single regulators to expanded chain of regulatory modules; (2) modules at adjacent phases often share common component that can bridge the continuity of the cycle; (3) there are modules at specific checkpoints (branchpoints) that allow cell entry or exit of the cycle according to external signals (Figure 4). Experimental verification is necessary to confirm any network predictions (Segal *et al.* 2003).

We are waiting for experimentalists to generate good quality data of ChIP-chip and expression from the same sample preparations for mammalian systems as well as to sequence multiple vertebrate genomes. Mammals alone are not enough for *cis*-element

studies about human; one needs distant organisms (such as chicken, for phylogenetic footprinting) as well as close ones (such as chimpanzee, for phylogenetic shadowing).

CONCLUSIONS

It is clear now that, having a “periodic table” of genes is not enough, we also need a network diagram telling us how the genes are connected and for this, we are going to need another “periodic table” of gene regulatory elements. Combination of computational and functional genomics will help us to filling up these tables quickly. Infrastructure such as promoter databases and *cis*-element/trans-factor databases is urgently needed. New technologies that can provide different genomewide view of the regulatory networks and new algorithms that integrate various large-scale data will be the keys for attacking human gene regulation problems (Banerjee and Zhang 2002). Conservation is important for revealing function; non-conservation can be even more important for understanding evolution (Wray *et al.* 2003). The recent discovery of a promoter that acquired p53 responsiveness during primate evolution through microsatellite expansion of weak binding sites (Contente *et al.* 2003) is an amazing testimony, and for this, one would have to look beyond just rodents.

Acknowledgements

I would like to thank all (including previous) members of Zhang lab and my collaborators for contributing most of the data and the figures, many before publications. Zhang lab is supported by grants (HG01696, GM60513, CA81152, CA88351) from NIH.

References

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., Harris M.A., Hill D.P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J.C., Richardson J.E., Ringwald M., Rubin G.M., and Sherlock G.. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25.

Antequera F. and Bird A. 1993. Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. USA* **90**:11995.

Bajic V.B., Seah S.H., Chong A., Zhang G., Koh J.L.Y., and Brusic V. 2002. Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoter. *Bioinformatics* **18**:198.

Bajic V.B. and Seah S.H. 2003. Dragon Gene Start Finder identifies approximate locations of the 5' ends of genes. *Nucl. Acid. Res.* **31**:3560.

Balija V., Nascimento L., Dike S., Zutavern T., Oui J., Palmer L., Hannon G., Xuan Z.Y., Zhang M.Q., and McCombie W.R. 2003. The mammalian gene set: systematic examination of gene predictions in mouse genome. Submitted.

Banerjee N. and Zhang M.Q. 2002. Functional genomics as applied to mapping transcription regulatory networks. *Current Opinion in Microbiology* **5**:313.

Banerjee N. and Zhang M.Q. 2003. Identifying cooperativity among transcription factors controlling yeast cell cycle. Submitted.

Beiley T.L. and Elkan C.P. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Intell. Sys. Mol. Biol.* **2**:28.

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

Bussemaker H.J., Li H. and Siggia E.D. 2001. Regulatory element detection using correlation with expression. *Nat Genet.* **27**:167.

Chen G.X. Hata N. and Zhang M.Q. 2003. Transcription factor binding element detection using functional clustering of mutant expression data. Submitted.

Chen J.J., Lee S., Zhou G., Rowley J.D. and Wang S.M. 2003. Generation of longer cDNA fragments from SAGE tags for gene identification. *Methods Mol. Biol.* **221**:207.

Chiang D.Y., Moses A.M., Kellis M., Lander E.S., and Eisen M. 2003. Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts. *Genome Biol.* **4**:R43.

Cliften P., Sudarsanam P., Desikan A., Fulton L., Fulton B., Majors J., Waterson R., Cohen B.A. and Johnston M. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* **301**:71.

Conlon E.M., Liu X.S. Lieb J.D. and Liu J.S. 2003. Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci U S A*. **100**:3339.

Contente A. Zischler H., Einspanier A., and Dobbelsstein M. 2003. A promoter that acquired p53 responsiveness during primate evolution. *Cancer Res*. **63**:1756.

Das M., Burge C.B., Park E., Colinas J., and Pelletier J. 2001. Assessment of the total number of human transcription units. *Genomics* **77**:71

Davuluri R.V., Suzuki Y., Sugano S., and Zhang M.Q. 2000. CART classification of 5'UTR sequences. *Genome Res*. **10**:1807.

Davuluri R.V., Grosse I., and Zhang M.Q. 2001. Computational identification of promoters and first exons in the human genome. *Nat. Genet*. **29**:412.

Davuluri R.V., Grosse I., and Zhang M.Q. 2003. Annotation of promoters and first exons in the human genome. Submitted.

Fickett J.W. and Hatzigeorgiou A.G. 1997. Eukaryotic promoter recognition. *Genome Res.* **7**:861.

Gardiner-Garden M. and Frommer M. 1987. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**:261.

GuhaThakurta D. and Stormo G.D. 2001. Identifying target sites for cooperatively binding factors. *Bioinformatics* **17**:608.

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

Hannenhalli S. and Levy S. 2001. Promoter prediction in the human genome. *Bioinformatics* **17** Suppl. 1:S90.

Hata N. and Zhang M.Q. 2003. BEAST: Tools for Transcription Factor Binding Site Search. Submitted.

Hertz G.Z., Hartzell G.W.^{3rd}, and Stormo G.D. 1990. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.* **6**:81.

Hubbard T., Barker D., Birney E., Cameron G., Chen Y., Clark L., Cox T., Cuff J., Curwen V., Down T., Durbin R., Eyraas E., Gilbert J., Hammond M., Huminiecki L., Kasprzyk A., Lehvaslaiho H., Lijnzaad P., Melsopp C., Mongin E., Pettett R., Pocock M., Potter S., Rust A., Schmidt E., Searle S., Slater G., Smith J., Spooner W., Stabenau A., Stalker J., Stupka E., Ureta-Vidal A., Vastrik I., and Clamp M. 2002. The Ensembl genome database project. *Nucl. Acid. Res.* 30:38.

Ince T.A. and Scotto K.W. 1995. A conserved downstream element defines a new class of RNA polymerase II promoters. *J. Mol. Chem.* **270**:30249.

Ioshikhes I.P. and Zhang M.Q. 2000. Large-scale human promoter mapping using CpG islands. *Nat. Genet.* **26**:61.

Kapranov P., Cawley S.E., Drenkow J., Bekiranov S. Strausberg R.L., Fodor S.P., and Gingeras T.R. 2003. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**:916.

Kato M., Hata N., and Zhang M.Q. 2003. Identification of cooperativity among transcription factors controlling cell cycle in yeast. Submitted.

Kel A.E., Kel-Margoulis O.V., Farnham P.J., Stephanie M.B., Wingender. E., and Zhang M.Q. 2001. Computer-assisted identification of cell cycle-related genes - new targets for E2F transcription factors. *J. Mol. Biol.* **309**:99.

Kellis M., Patterson N., Endrizzi M., Birren B., and Lander E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**:241.

Korf I., Flicek P., Duan D., and Brent M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17** Suppl.11:S140.

Levine M. and Tjian R. 2003. Transcription regulation and animal diversity. *Nature* **424**:147.

Lawrence C.E., Altschul S.F., Bogouski M.S., Liu J.S., Neuwald, A.F. and Wooten J.C. 1993. Detecting subtle sequence signals: A Gibbs sampler strategy for multiple alignment. *Science* **262**:208.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA. 2002. *Transcriptional regulatory networks in Saccharomyces cerevisiae.* *Science* **298**:799.

Li Z., van Calcar S., Qu C., Kolodner R., Cavenee W.K., Zhang M.Q. and Ren B. 2003. A global transcriptional Regulatory role for c-myc in Burkitt's lymphoma cells. *Proc. Natl. Acad. Sci. USA* **100**:8164.

Liu X.S., Brutlag D.L. and Liu J.S. 2002. An algorithm for finding protein-DNA binding sites with application to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.* **20**:835.

Neuwald A.F., Liu J.S., and Lawrence C.E. 1995. Gibbs motif sampling: Detection of bacterial outer membrane repeats. *Protein Sci.* **4**:1618.

Pennacchio L.A. and Rubin E.M. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat Rev. Genet.* **2**:100.

Ponger L. and Mouchiroud D. 2002. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics* **18**:631.

Prestridge D.S. 1995. Predicting Pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.* **249**:923.

Ren B., Cam H., Takahashi Y., Volkert T., Terragni J., Young R.A. and Dynlacht B.D. 2002. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev.* **16**:245.

Scherf M., Klingenhoff A., and Werner T. (2000). Highly specific location of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.* **297**:599.

Segal E., Shapira M., Regev A., Pe'er D., Botstein D, Koller D., and Friedman N. 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**:166.

Smale S.T. and Kadonaga J.T. 2003. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **72**:449.

Solovyev V.V. 2002. Finding genes by computer: probabilistic and discriminative approaches. In *Current Topics in Computational Molecular Biology* (ed. T. Jiang *et al.*). p.201. The MIT Press. Cambridge, Massachusetts.

Suzuki, Y., Ishihara, D., Sasaki, M., Nakagawa, H., Hata, H, Tsunoda, T., Watanabe, M., Komatsu, T., Ota, T., and Isogai, T. 2000. Statistical analysis of the 5' untranslated region of human mRNA using "Oligo-Capped" cDNA libraries. *Genomics* **64**: 286.

Takai D. and Jones P.A. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA* **99**:3740.

Tasic B., Nabholz C.E., Baldwin K.K., Kim Y., Rueckert E.H., Ribich S.A. Cramer P., Wu Q., Axel R., and Maniatis T. 2002. Promoter choice determines splice site selection in protocadherin alpha and gamma pre-mRNA splicing. *Mol Cell*. **10**:21.

Trinklein N.D., Aldred S.J., Saldanha A.J., and Myers R.M. 2003. Identification and functional analysis of human transcriptional promoters. *Genome Res*. **13**:308.

van Helden J. Rios A.F., and Collado-Vides J. 2000. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucl. Acid. Res*. 28:1808.

Wang X., Su H. and Bradley A. 2002. Molecular mechanisms governing Pcdh-gamma gene expression: evidence for a multiple promoter and cis-alternative splicing model. *Genes Dev*. **16**:1890.

Weinmann A.S., Yan P.S., Oberley M.J., Huang T.H., and Farnham P.J. 2002. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev*. **16**:235.

Weinmann A.S., Bartley S.M., Zhang T., Zhang M.Q., and Farnham P.J. 2001. The Use of Chromatin Immunoprecipitation to Clone Novel E2F Target Promoters. *MCB* **21**:6820.

Wray G.A., Hahn M.W., Abouheif E, Balhoff J.P., Pizer M., Rockman M.V., and Romano L. 2003. The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* In press [Epub ahead of print].

Wu Q., Zhang T., Cheng J.-F., Kim Y., Grimwood J., Schmulz J., Dickson M., Noonan J.P., Zhang M.Q., Myers R.M. and Maniatis T. 2001. Comparative DNA sequence analysis of mouse and human protocadherin gene clusters. *Genome Res.* **11**:389.

Xuan Z.Y., Wang J.H., and Zhang M.Q. 2002. Computational comparison of two mouse draft genomes and the human goldenpath. *Genome Biology* **4**:R1

Yeh R.F., Lim L.P., and Burge C.B. 2001. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11**:803.

Zhang M.Q. 1998a. A Discrimination Study of Human Core-promoters, in *Proceedings of Pacific Symposium on Biocomputing 1998* (ed. R.B. Altman et al.), p.240, World Scientific, Singapore.

Zhang M.Q. 1998b. Identification of human gene core promoters in silico. *Genome Res.* **8**:319.

Zhang M.Q. 1999a. Large scale gene expression data analysis: a new challenge to computational biologists. *Genome Res.* **9**:681.

Zhang M.Q. 1999b. Promoter Analysis of Co-regulated Genes in the Yeast Genome. *Computers and Chemistry* **23**:233.

Zhang M.Q. 2000a. Discriminant analysis and its application in DNA sequence motif recognition. *Briefings in Bioinformatics* **1**:331.

Zhang M.Q. 2002a. Computational Prediction of Eukaryotic Protein Coding Genes. *Nat. Rev. Genet.* **3**:698.

Zhang M.Q. 2002b. Computational methods for promoter recognition. In *Current Topics in Computational Molecular Biology* (ed. T. Jiang *et al.*). p.201. The MIT Press. Cambridge, Massachusetts.

Zhang T. and Zhang M.Q. 2001. Promoter Extraction from GenBank (PEG): automatic extraction of eukaryotic promoter sequences in large sets of genes. *Bioinformatics* **17**:1232.

Zhu J. and Zhang M.Q. 1999a. SCPD: A Promoter Database of Yeast *Saccharomyces cerevisiae*, *Bioinformatics* **15**:607.

Zhu J. and Zhang M.Q. 2000. Cluster, function and promoter: analysis of yeast expression array. *Pac. Symp. Biocomput.* 479.

Figure legends

Figure 1. Flow chat for CSHL Mammalian Promoter Database.

Figure 2. Demonstration of “analysis on-the-fly” utility: a user can paste in a list of genes (accession numbers) and specify the range (-700,+300), extract promoter sequences (including orthologous sequences), select P53 gene promoters, do promoter alignment and identify the motif in conserved regions.

Figure 3. Luciferase activity of predicted promoters in a 120kb region of human ch19q13 (provided by Lisa Stubbs).

Figure 4. Reconstructed yeast cell cycle transcriptional regulation network (adapted from Kato *et al.* 2003).

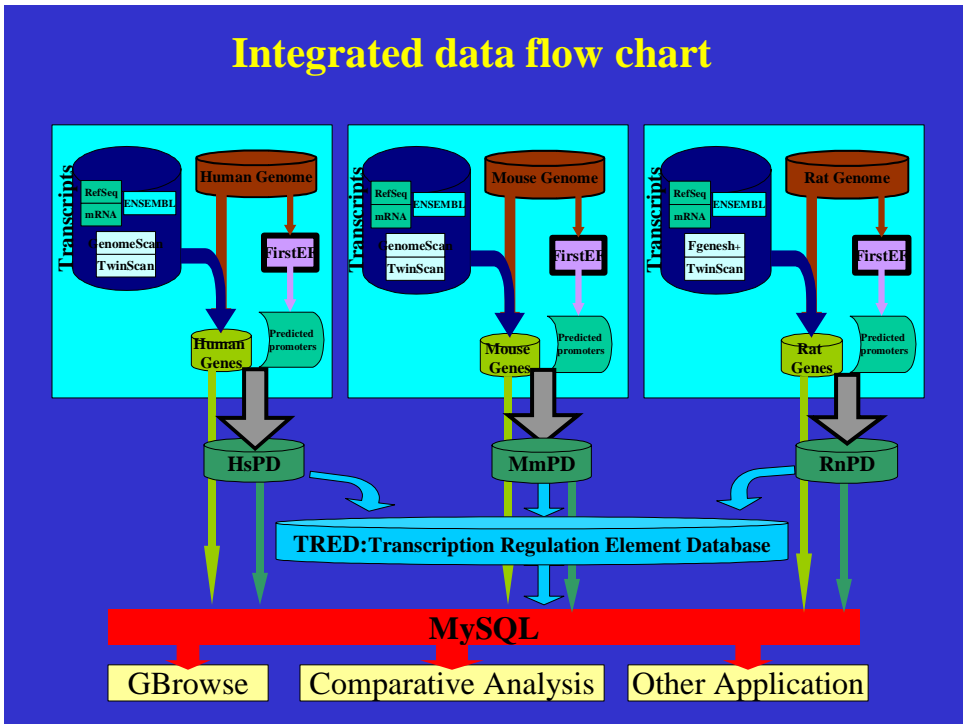


Figure 1

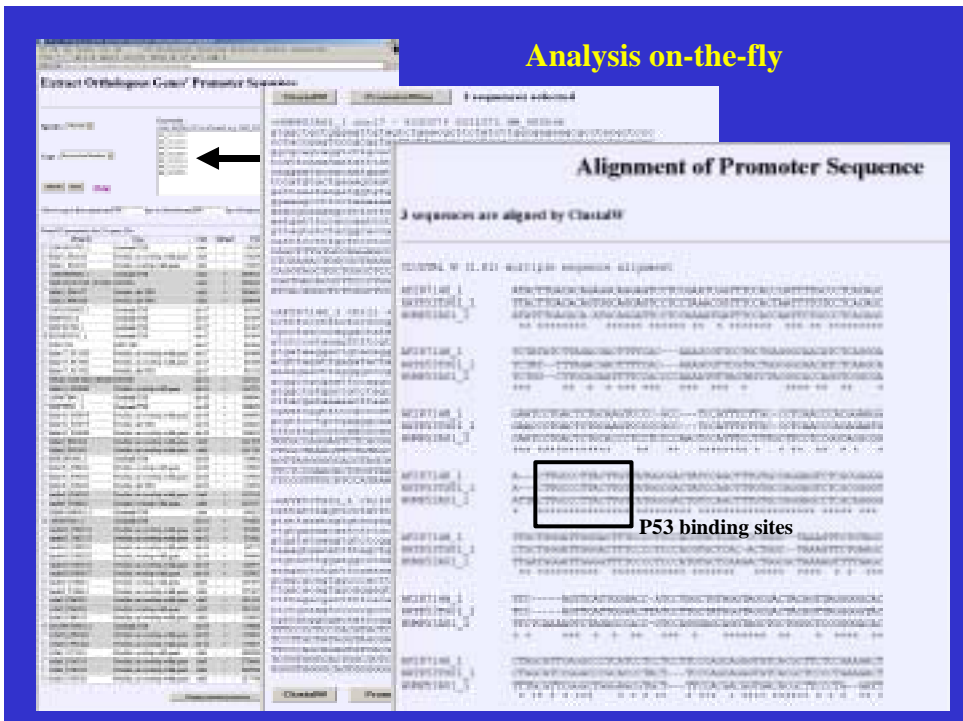


Figure 2

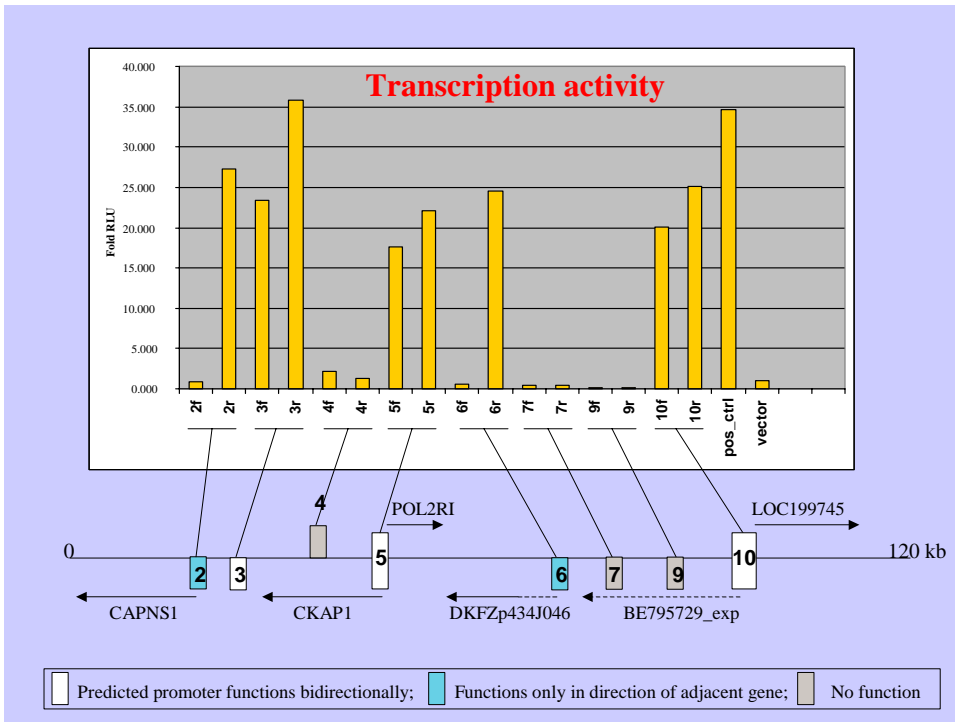


Figure 3

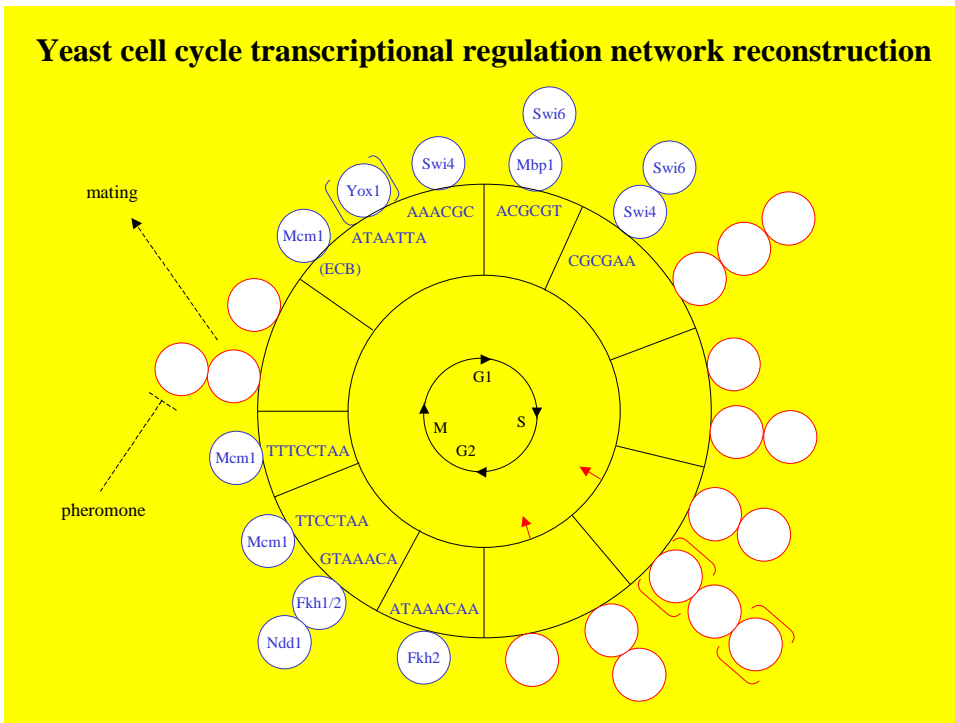


Figure 4