

SMASHing regulatory sites in DNA by human-mouse sequence comparisons

Mihaela Zavolan^{†1}, Nikolaus Rajewsky^{‡1}, Nicholas D. Socci^{*¶} and Terry Gaasterland[†]

[†]Laboratory for Computational Genomics, and

^{*}Laboratory for Molecular Genetics

The Rockefeller University, New York, NY 10021

[‡]Department of Biology

New York University, New York, NY 10003-6688

[¶]Department of Pathology

Seaver Foundation Program in Bioinformatics

Albert Einstein College of Medicine, Bronx, NY 10461

Running title: Regulatory sites from human-mouse comparisons

Keywords: Transcriptional regulation, binding sites, cross-species comparisons

For correspondence: M . Z. mihaela@genomes.rockefeller.edu and N. R. nikolaus.rajewsky@nyu.edu

Abstract

Regulatory sequence elements provide important clues to understanding and predicting gene expression. Although the binding sites for hundreds of transcription factors are known, there has been no systematic attempt to incorporate this information in the annotation of the human genome. Cross species sequence comparisons are critical to a meaningful annotation of regulatory elements since they generally reside in conserved non-coding regions. To take advantage of the recently completed drafts of the mouse and human genomes for annotating transcription factor binding sites, we developed SMASH, a computational pipeline that identifies thousands of orthologous human/mouse proteins, maps them to genomic sequences, extracts and compares upstream regions and annotates putative regulatory elements in conserved, non-coding, upstream regions. Our current dataset consists of approximately 2500 human/mouse gene pairs. Transcription start sites were estimated by mapping quasi-full length cDNA sequences. SMASH uses a novel probabilistic method to identify putative conserved binding sites that takes into account the competition between transcription factors for binding DNA. SMASH presents the results via a genome browser web interface which displays the predicted regulatory information together with the current annotations for the human genome. Our results are validated by comparison to previously published experimental data. SMASH results compare favorably to other existing computational approaches.

1 Introduction

Genome annotation has so far been limited primarily to the identification and functional assignment of coding regions (genes). Yet the genome also contains abundant sequence signals for gene regulation, which provide important clues to predicting and interpreting the expression patterns of genes. In particular, transcription factors recognize and bind certain sequence elements (binding sites, sequence motifs) in non-coding regions.

¹These authors contributed equally to the work.

Their presence at a given genomic locus (cis-regulatory sequence) controls the rate of transcription of proximal genes. For vertebrates, hundreds of transcription factors are known, and the cognate motifs of roughly 150 factors have been estimated by Wingender et al. (2001) from thousands of available instances. Cross-species comparisons are key for predicting functional binding sites since conserved non-coding regions are highly likely to contain the regulatory sites (Hardison et al., 1997; Wasserman et al., 2000; Loots et al., 2000). With the availability of genomic sequences for multiple eukaryotic species, it becomes feasible to start annotating non-coding sequences in terms of binding sites for known transcription factors.

Here we describe SMASH (Sequence Motif Annotation based on Sequence Homology), a fully automated procedure which operates on protein, cDNA and genomic sequences, and on sequence motifs for known transcription factors. We apply SMASH to human and mouse data. Using the RefSeq database of human and mouse protein and cDNA sequences, the public assemblies of the human and mouse genomes, and the TRANSFAC database of transcription factor binding sites, SMASH

- identifies thousands of pairs of orthologous human and mouse genes,
- maps them unambiguously to the corresponding genome,
- estimates transcription starts,
- identifies conserved non-coding regions,
- predicts binding sites for known transcription factors in conserved non-coding regions proximal to transcription start.
- generates annotation information which is integrated into a genome annotation system available over the web (<http://hal.rockefeller.edu/Smash>).

The entire system serves as a starting point for genome-wide characterization of gene regulatory interactions. Our methods differ from existing approaches (Loots et al. (2002), Jegga et al. (2002), Dieterich et al. (2002), Boris Lenhard, personal communication) by several key aspects. We have automated the entire procedure, from inferring orthologous proteins to reporting upstream regulatory elements. We attempt to automatically identify transcription start sites from full length cDNA's. Our method of extracting conserved sequences is flexible and can accept output from any of a number of local alignment algorithms; in particular, a global alignment is not needed. We have implemented a novel probabilistic method of *simultaneously* searching for binding sites for several transcription factors. The algorithm models the *competition* of several transcription factors (weight matrices) for a stretch of DNA. This approach should better reflect the biochemistry of protein-DNA interactions than approaches treating the factors independently. Finally, we have incorporated and made available the results via the web using the Generic Genome Browser, which integrates our output the currently available public annotation for the human genome.

Characterization of regulatory sequences in the genomes of vertebrates is a complex task for the following reasons. The length of regions in which regulatory motifs are usually found is large, typically tens of kilobases per gene, while the length of transcription factor binding sites is small, on the order of a dozen bases. Combined with the “fuzzy” (degenerate) nature of most binding sites this usually leads to a very high number of false positives. To enhance the likelihood of identifying functional binding sites, we focused the search on non-coding regions which are conserved between human and mouse, and we used a species-specific model for scoring regions which are not part of binding sites (background sequence).

A gene is typically regulated by transcription factors binding within a range of tens of kilobases upstream and downstream from the transcription start (promoter). Since the transcription start sites of genes are difficult to identify, and can be located far away from the more readily identifiable translation start, we restrict our analysis to genes for which a sufficient length of 5' untranslated region is contained in the cDNA. This enhances the likelihood of correctly identifying the first (possibly non-coding) exon and, therefore, transcription start. With this approach we have estimated transcription start sites for more than 2500 genes. Several large scale projects for sequencing full length cDNA are under way, and data for a large fraction of the remainder of human/mouse genes will be available in the near future. Thus, it will be possible to analyze the majority of human/mouse genes using this method.

Finally, and perhaps most importantly, transcriptional regulation in multicellular eukaryotes is thought to be largely combinatorial: multiple transcription factors frequently bind in spatial proximity, and the rate of transcription of the gene they control depends on the concentrations of the different factors (Xanthopoulos et al., 1991; Arnone et al., 1997; Ptashne and Gann, 1999). Rather than scanning the conserved non-coding regions with one weight matrix at a time, we base our predictions on a probabilistic *simultaneous* fit of conserved sequences to sets of weight matrices. The weight matrix (Berg and von Hippel, 1987) is a surrogate for the energetic preference of a transcription factor for a particular sequence, and our algorithm models the

competition of several factors and their binding energies for a stretch of DNA. We furthermore used non-conserved regions to build an empirical model of the compositional biases expected in non-coding DNA, which we specifically take into account in scoring.

2 Results

2.1 Extraction of orthologous gene pairs

Starting with 15,310 human and 8,578 mouse proteins from the RefSeq (June 2002 release), we performed pairwise protein sequence comparisons using the BLAST algorithm, and extracted pairs of sequences which were each other's best match, with each member of the pair being covered at least 70% of its length by matches. This procedure identified 5,346 putative orthologous protein pairs.

2.2 Transcription start site approximation

The review process behind the RefSeq database includes sequence extension when more data becomes available. However, cDNA sequences deposited in Genbank are rarely full-length. 9% of RefSeq sequences have no 5'UTR. Large-scale efforts to trap and sequence full-length mRNA sequences are currently under way (Suzuki et al., 1997; Riken Phase II Team and Fantom Consortium, 2001). When we compare the length of the 5'UTR between RefSeq sequences and a set of 954 full-length human cDNA sequences (Suzuki et al., 2000), we find that a higher proportion of RefSeq sequences have UTR lengths lower than 55-60 nucleotides. This indicates that such RefSeq sequences are most probably incomplete and we decided to only use sequence pairs in which at least one of the human or mouse cDNA's has a 5'UTR length of at least 55 nucleotides. This reduces our dataset to 4,380 pairs of orthologous genes. Stringent mapping of these pairs (see 4.2) to the corresponding genomes yields 2,577 pairs of orthologous genes for which we analyzed upstream regions. Our mapping of the human RefSeq sequences was similar to the mapping obtained by the Human Genome Project (<http://genome.ucsc.edu>). Inspection of the annotations of these orthologous pairs indicates that they are indeed true orthologs. To anchor the transcription start site in both species, we extracted 100kb upstream of the translation start of each of the two orthologous genes, and we aligned these upstream sequences to each other using the BLAST algorithm (section 4.2). The 5' untranslated regions map somewhere within the 100kb region upstream of translation start. We used the genome mapping coordinates of the 5' untranslated regions to identify the 5'-most cDNA nucleotide aligned from either species. This position was considered to be the transcription start site.

2.3 Extraction of conserved non-coding regions

Upstream regions were aligned using the BLAST algorithm with a permissive scoring scheme (word size $W = 7$, mismatch penalty $q = -2$) to extract regions with lower conservation level than that expected in the coding regions (for which BLAST has been optimized). We then chained the local alignments as described in section 4.3. This chaining resolves conflicts between overlapping local alignments and defines blocks of conserved, syntenic genomic regions from which we extracted conserved non-coding regions. These are genomic regions conserved between human and mouse, situated between the apparent transcription start and either the start of another gene, or the start of a gap in the assembly of the human genome. The annotation of the human genome using RefSeq transcripts, provided by the Human Genome Sequencing Consortium, was used as the basis for identifying the location of upstream genes.

90% of the conserved non-coding regions are between 19 and 240 nucleotides long. The percent identity in these regions ranges between 44% and 100%, with a median of 85%. The median coverage of upstream non-coding regions by conserved blocks is 7% of the length of the sequence (and the mean is 16%). This value is comparable to that reported by Jareborg et al. (1999) using a different method, on a smaller dataset.

2.4 Identification of putative transcription factor binding sites

We used a set of 32 weight matrices from the TRANSFAC database (Wingender et al., 2001) including CEBPB, CMYB, CREB, E2F, ER, Elk1, GR, HNF1a, HNF3b, HNF4, NFAT, NF κ B, P53, SP1, SREBP1, SRF, and STATx. These matrices were chosen based primarily on their quality and the amenability to experimental testing. We also extracted the sequences between conserved blocks and we constructed empirical models for the background, non-conserved human and mouse sequence. We applied our novel scoring method (section 4.5) to conserved non-coding sequences to predict putative binding sites and background regions. This method

takes into account the competition between transcription factors that bind to similar, overlapping sequence motifs in the genome. In contrast to other methods that score each position in the genome independently, our method has the virtue that it captures some aspects of the biochemistry of DNA-protein interactions. These sites, together with information about their score and position relative to the transcription and translation start in the human genome, are available over the web at <http://hal.rockefeller.edu/Smash>.

2.5 Validation of predicted binding sites

To validate our method for extracting conserved sequences and identifying binding sites we performed three different tests (see below). The first test demonstrates that binding sites predicted by SMASH correlate significantly with the experimentally known regulation of a test set of 11 genes. The data used in this test did not include any information about the position of the binding sites in the genome. The second test indicates that SMASH recovers the position of known functional binding sites with good specificity (fraction of true positives among predictions). The third test shows that our novel probabilistic method of predicting binding sites which takes into account the competition between factors, allows us to achieve a higher specificity than other existing approaches.

2.5.1 First Test

We selected from the gene list in section 2.4 a test set of 11 genes which have been previously shown experimentally to be regulated by a total of 14 factors (Table 1). These include interferon-regulated factor 1, hepatic nuclear factor 4, JunB, phosphoenol pyruvate carboxykinase, DNA damage-inducible gene GADD45, and TNF receptor-associated factor. Then, for various probability levels we extracted all binding sites predicted by SMASH and counted how many of these sites correspond to factors known to regulate the genes (Table 2). For example, at a probability level of 0.9 we predict 20 sites corresponding to factors known to regulate our 11 genes. These represent 36% of all predicted inputs for these genes (p-value=0.025). The number of true positives is likely to be even higher because the experimental data are certainly not exhaustive. We computed the significance of our results by sampling 1000 random subsets of 11 genes from our total dataset and calculating the probability to obtain the same or better fraction of true positives among the total number of predicted sites (R values in table 2) by chance (p-value). All our R values are significant at a level of at least 0.05.

Table 1: Gene names and GI numbers with associated factors (dataset for test 1 (section 2.5.1)).

Gene name	GI number	Factors
Interferon-regulatory factor 1	4504720	SP1, NF κ B, STAT
Hepatic nuclear factor 4	4504442	HNF, CEBP, GR
Jun B	4504808	ELK 1, SRF, STAT, NF-Y, SP1
Cytokine-inducible SH2-containing protein	19923410	STAT
Phosphoenol pyruvate carboxy kinase	4505638	HNF
Serum/GR-regulated kinase	20127540	SRF, GR
Minichromosome maintenance deficient 5	6981191	E2F
Growth arrest and DNA damage-inducible alpha	9790904	OCT1, p53
Growth arrest and DNA damage-inducible beta	9945331	STAT
TNF receptor associated factor 1	5032192	NF κ B

^a See references: Zhao et al. (2000); Chin et al. (1997); Zhan et al. (1993); Kastan et al. (1992) ^b Jin et al. (2001); Takahashi et al. (2001) ^c Cassuto et al. (1997); Yanuka-Kashles et al. (1994); O'Brien et al. (1995) ^d Schwenzer et al. (1999); Wang et al. (1998); De Smaele et al. (2001) ^e Ohtani et al. (1999); Ishida et al. (2001); Leone et al. (1998) ^f Coffey et al. (1995) ^g Hipskind et al. (1994); Kitabayashi et al. (1993) ^h Abdollahi et al. (1991); Thompson et al. (2000) ⁱ Sims et al. (1993); Yang et al. (2002); Rein et al. (1994) ^j Kraus et al. (1999) ^k Starr et al. (1997); Matsumoto et al. (1997); Mui et al. (1996) ^l Bailly et al. (2001)

2.5.2 Second test

As an additional test we analyzed a set consisting of orthologous regulatory sequences for 14 human/mouse genes for which 40 functional binding sites have been defined experimentally: skeletal muscle actin (Genbank accession: AF182035), aldolase (X12447), alpha B crystallin (M28638), cardiac myosin heavy chain (U71441), cEBP alpha (U34070), cdc2 (L06298), cholesterol 7- α -hydroxylase (L13460), early growth response protein 1 (AJ243425), glucose-6-phosphatase (AF051355), leptin (U43589), lipoprotein lipase (M29549), creatin kinase (M21487), retinoblastoma susceptibility gene (L11910), troponin I (L21905). This dataset together with weight-matrices for a total of 14 factors was generously provided to us by Boris Lenhard et al (submitted). The total sequence length analyzed is 49006 bases (roughly 25 K per species). The level of conservation, as defined by our method, is relatively high (35%) compared to our average conservation of 16%. The reason for this is that most of the test sequences come from locations close to the 5' end of genes, where the level of conservation is generally higher (data not shown). We found that 100% of the experimental sites reside in conserved regions. We ran our probabilistic method of predicting binding sites (section 4.5) on (a) all single sequences which contain at least one experimental site and (b) all conserved sequence pairs. As before, we estimate the quality of our predictions in terms of the R , the fraction of correctly identified experimental sites among all predicted site. The results are shown in Table 3. Our R values are larger by a factor of 2 – 5 compared to Lenhard et al (submitted).

For example, at a level of 0.1 probability for individual predicted binding sites we recover 23 (18) or 58% (45%) of the experimental sites while predicting 4664 (1229) sites in single sequences (conserved sequences) ($R = 0.004, 0.015$). At this very high rate of recovery of known sites, on the average all of the sequences are covered with predicted binding sites. Interestingly, not all of the 40 experimental sites are recovered which may be due to the accuracy with which weight matrices reflect the sequence motifs to which the factors bind. 10 more binding sites could be recovered when substituting some of the weight matrices which did not recover any experimental sites with weight matrices from TRANSFAC. Thus, a better definition of the weigh matrices is likely to improve the rate at which true sites are recovered.

2.5.3 Third test

Following Loots et al. (2002), we used SMASH to analyze the cytokine gene cluster containing the genes for IL-3, IL-4, IL-5, IL-13 and GMCSF genes (approximately 1 megabases in each species). We were able to map 15 known NFAT binding sites (Loots et al., 2002) to the current assemblies of the human and mouse genomes. We found that although only 10% of the whole locus was conserved, 13 of the NFAT sites are located in conserved regions, a result which is similar to that reported by Loots et al. (2002). The ratio of the number of NFAT sites recovered by SMASH versus the total number of predicted NFAT sites is smaller than that in Loots et al. (2002). For example, to predict 11 out of the 13 binding sites we make 266 predictions (4.1% true positives), while 734 predictions are made in Loots et al. (2002) (1.5% true positives). This gives us 2.8 reduction in the rate of false positives. Interestingly, if we require that the predicted sites have at least 0.7 probability, we can further improve the sensitivity with which we recover experimental sites: we find 7 of the 13 known site, while making only 58 predictions. The difference seems largely due to the competition between the transcription factors in our method because if we eliminate almost all of the 32 factors from our dataset and keep only those used in (Loots et al., 2002), we arrive at results very similar to (Loots et al., 2002).

Probability threshold	Total predictions	With functional evidence	Ratio	p value
0.1	1451	220	0.15	< 0.01
0.6	238	60	0.25	< 0.01
0.75	136	45	0.33	< 0.005
0.9	55	20	0.36	< 0.025
0.95	19	7	0.37	< 0.05

Table 2: SMASH performance (Sensitivity as a function of detection threshold) on genes with known regulation.

Probability	Single sequence			Conserved sequence		
	total	true	R	total	true	R
0.10	4664	23	0.004	1229	18	0.015
0.60	670	8	0.012	191	7	0.037
0.75	337	6	0.018	106	7	0.067
0.90	115	2	0.017	39	2	0.051
0.95	50	2	0.040	25	2	0.080

Table 3: Performance for both single sequence and conserved sequence algorithms (test 2, see 2.5.2).

2.6 Visualization tools

Although there are a number of browser systems which display a variety of information about the human and mouse genome none of them offers transcription factor binding site annotations. And while there are several works examining regulatory sequences they do not allow to view this information together with the other genome annotations. To set up such a system we have adapted Lincoln Stein’s Generic Genome Browser (<http://www.gmod.org>) to display our cDNA mappings, conserved non-coding regions and computed bindings sites superimposed on the publicly available annotation of the human genome. The results can be viewed at <http://hal.rockefeller.edu/Smash>. Figure 1 shows a screenshot with the SMASH annotation of the human interferon regulatory factor 1 locus. This website will be continually updated as more cDNA sequence or annotation is made available and improved binding site scoring methods are developed. Finally, our browser is the starting point for a web-based system, which will allow users to input their own cDNA mapping and/or transcription factor information to view their results in the context of available annotations.

3 Discussion

Our ultimate goal is to develop SMASH into an interactive web-based tool where the user can specify the genes and weight matrices (transcription factors) of interest to obtain predicted transcription factor binding sites. The current paper presents the methods which we developed for this purpose. Special care was taken to design a pipeline which (a) is not dependent on the specific local alignment method for detecting conserved sequences, (b) is capable of identifying transcription starts by using sufficiently long cDNA’s, (c) takes the competition between transcription factors into account, (d) incorporates the results into a standard annotation system, accessible via a web browser. None of these features is available in other approaches. We have demonstrated that for the prediction of binding sites, competition between transcription factors (c) results in much higher rate of true positives when compared to existing approaches (section 2.5).

The number of genes which we can analyze (more than 2500 right now) will grow rapidly in the near future as the number of available full length cDNA will grow. A number of large scale full length cDNA sequencing projects are under way (Riken Phase II Team and Fantom Consortium, 2001; Suzuki et al., 2000). It is also possible to use recent computational methods to detect promoter start sites (for example Davuluri et al. (2001)) to increase the size of our dataset.

Our method of predicting binding sites assumes the presence of each factor in the cell. However, it is straightforward to include information about the concentrations of the factors in the algorithm (Rajewsky & Zavolan, in preparation). It would be most interesting to use microarray data to estimate transcription factor concentrations in order to predict binding sites and thus regulatory inputs which are active under the conditions of the experiment. Several other extensions are possible, for example one could use a probabilistic algorithm to correlate the presence of sequence motifs with whole-genome expression data. These sequence motifs could either be our annotated sites, or novel sequence motifs. Several studies have demonstrated the power of such approaches (Holmes and Bruno, 2000; Bussemaker et al., 2001); however, no study so far has also integrated cross-species comparisons.

We have used Lincoln Stein’s Generic Genome Browser as a framework for integrating our predictions with curated or experimental annotation. This integration is extremely useful in assessing the validity and relevance of the computational predictions. This system can be extended with new or updated information as it becomes available and investigators can copy and customize it for their specific needs. In the future we will add the interface necessary for user to input their own weight-matrices and interactively search for binding sites.

It may be the case that for some factors the binding site is not well defined when considered alone but only when it is placed in the context of surrounding binding sites. Obviously, SMASH can already be used to look for the co-occurrence of certain conserved sites; however, it is also possible to use a probabilistic method similar to that of Rajewsky et al. (2002) which searches directly for clusters of binding sites.

4 Methods

4.1 Identification of orthologous genes

Our protein sequence dataset consisted of the June 2002 release of the RefSeq database (Pruitt and Maglott, 2001). We reasoned that a preliminary computational analysis is required for promoting a sequence to RefSeq provisional status, thus ensuring a minimal quality of the sequence data.

We used a sequence-based approach to identify orthologous genes. The mouse and human protein datasets were reciprocally compared using the `blastp` program (Altschul et al., 1997) (using default parameters). For each sequence, we extracted the best match in the other set and calculated the fraction of the protein length involved in perfect matches with its best-matching partner. Pairs of proteins which were each other's best match, with at least 70% of both sequences being identically mapped, were considered to be orthologous.

4.2 Genome mapping and approximate identification of transcription start

By mapping full-length cDNA sequences to the genome, we can improve the accuracy with which regions upstream of the transcription start sites are identified. This is especially important for us since most of the regulatory elements involved in transcription regulation are found within kilobases (kb) of the start of transcription. However, full-length cDNA sequences are only recently starting to become available, and the cDNA sequences in GenBank have various degrees of completeness of the 5' UTR. We found that 5'UTR lengths shorter than 55-60 nucleotides are more frequent among RefSeq sequences than among human full-length cDNAs. Therefore, to approximate transcription starts from cDNA that may be incomplete, we require that for each orthologous gene pair at least one cDNAs has a minimum of 55 nucleotides of 5' untranslated region (UTR). If the gene does indeed have a first non-coding exon, this UTR sequence allows us to anchor the first exon to the genome.

We restricted ourselves to pairs of orthologs for which we have a correct and relatively complete mapping to the genome in both species. This is because both the human and the mouse genome are not finished and far from perfect. To speed up the mapping CDS and cDNA sequences were mapped to the Santa Cruz assembly of the human genome (April 2002 freeze) and the Whitehead assembly of the mouse genome, using a combination of Megablast, BLAST and chaining. CDS sequences were mapped using Megablast, and the local sequence alignments were chained, as described below, to extract approximate gene structures. For a CDS sequence to be considered well mapped, over 90% of its length had to be involved in perfect matches to the genome. Sequences which did not pass this filter were submitted to a more sensitive alignment, using BLAST followed by chaining. The cDNA sequences were then aligned to the corresponding genomic loci using a similar strategy. To ensure accurate identification of transcription and translation start sites, we required that the cDNA alignment starts within at most 10 nucleotides from the start of the cDNA sequence, and that the translation start site is found within the alignment of the cDNA to the genome. The resulting set of 2,577 orthologs was used in the analysis of upstream regions.

4.3 Chaining together regions of high similarity

The problem of "chaining" together regions of good local alignment has been discussed in detail by a number of authors (Chao and Miller, 1995; Gusfield, 1997). We have implemented a version of the chaining algorithm described in Gusfield (1997), as described below. Assume that we are given strings S_1 and S_2 which we want to align to each other using some scoring scheme. Given any alignment, we can calculate its score under the scoring scheme. The task is to find the alignment that maximizes the score. In some cases, such as cDNA-to-genome or upstream region alignment, we are not interested in a global alignment of the two strings. Within bounds, we do not believe that very short introns are more likely than long introns, or short insertion elements more likely than longer ones. The kind of alignment algorithm that we want for these situations identifies regions of good local alignment (for example exons), and combines them into a chain, such that the relative position of these regions is identical between the two strings (for example genome and cDNA), and the combined score of this chain is maximal. The score for the chain is simply the sum of the scores of local alignments which are part of the chain.

In our implementation chaining is done independently of the search for regions of good local alignment. This allows us to use different modules for identifying these regions. For the purpose of this study, we used the BLAST algorithm (Altschul et al., 1997) to obtain local alignments between sequence pairs. In BLAST terminology, regions of good local alignment are called high scoring pairs (HSP). We use BLAST to extract the set of high scoring pairs for strings S_1 and S_2 . We then sort the HSP by orientation—we only allow fragments that map in the same orientation to be chained together—and by their start position in one of the strings. For each pair of HSP i and j , we determine whether HSP j can follow HSP i in the chain. Let $b_i^{S_1}$ and $b_i^{S_2}$ be the start coordinates of HSP i in S_1 and S_2 respectively. Because we want to preserve the local order of HSP between the two strings, HSP j can follow HSP i in the chain if $b_i^{S_1} \leq b_j^{S_1} \& b_i^{S_2} \leq b_j^{S_2}$. In this case, we say that i is a parent of j and we calculate the contribution of HSP j to the score, if

1. j were to follow i in the chain
2. j were to be the last HSP in the chain. We denote the contribution of HSP j to the score when it follows HSP i by c_{ji} . When HSP i and j do not overlap, this score reduces to the score of HSP j as given by the local alignment algorithm. Denote this score by γ_j .

Finally, we denote the set of parents of node i by P_i . We traverse the ordered list of HSP, and, for each HSP i , we calculate the maximum score that could be obtained in a chain that ends with HSP i (μ_i): if i does not have any parents, then $\mu_i = \gamma_i$, otherwise $\mu_i = \max_{j \in P_i} \mu_j + c_{ij}$. To get the chain, we trace back from the HSP with the highest score.

4.4 Extraction of conserved non-coding regions

We used the translation start to anchor the alignments, and then identified the putative transcription start as the pair of nucleotides situated farthest upstream of translation start, with one of the nucleotides in the pair being part of the human or mouse cDNA sequence. Regions of good local alignment were obtained using BLAST, with a parameter setting that allows identification of short regions of moderate similarity (wordsize $W = 7$ and mismatch penalty $q = -2$). The local alignments were chained as described in section 4.3.

RefSeq sequence-to-genome mappings were used to identify coding regions in the human genome. Only conserved upstream regions with no evidence of containing coding sequences were searched for transcription factor binding sites.

4.5 Identification of putative transcription factor binding sites

We devised a novel algorithm to parse conserved upstream sequence (pairwise sequence alignments) into putative binding sites and background sequence. We assume that the binding motifs are known. The output consists of probabilities for each of the transcription factors in the input set to bind at every position in the sequence. Following the BLAST terminology we call the local sequence alignments HSPs (high scoring pairs). We work with a weight matrix model for binding sites: for each position i of the binding site there is a fixed probability $p_i(b)$ to observe nucleotide b (Berg and von Hippel, 1987). The p_i 's can be computed (including pseudocounts) from a sample of N known binding sites simply by counting the number of occurrences $n_i(b)$ of each base b at each position i in the alignment of binding sites via $p_i(b) = \frac{n_i(b)+1}{N+4}$. As a first approximation, we regard every position in an HSP as either being part of a binding site or not. In the latter case we refer to the nucleotide at that position as being generated by a background model. We had sufficient non-conserved, inter-HSP sequence to construct a 6th order Markov model for the background. The use of a higher-order background models has been shown to improve the detection of binding sites (Thijs et al., 2001). We experimented with higher order (than 6th) models and found no substantial changes in our results.

Our goal is to parse the HSP's into binding sites and background using a scoring scheme that allows us to assign to each binding site a probability that its corresponding factor will indeed bind at that position. Let \mathbf{W} be the set of weight matrices for the transcription factors that we consider with their associated probabilities p^{W_j} and lengths $\sigma(W_j)$. Let us consider an HSP composed of human/mouse sequences s^H and s^M , respectively. The likelihood that the subsequence of length $\sigma(W_j)$ starting at position i in the human sequence,

$$s^H [i \dots (i + \sigma(W_j) - 1)] \quad (1)$$

(ie, a string from position i to $i + \sigma(W_j) - 1$) is a binding site for transcription factor j is

$$P(s^H[i \dots (i + \sigma(W_j) - 1)] | W_j) = \prod_{k=i}^{i+\sigma(W_j)-1} p_{k-i}^{W_j}(s^H[k]) \quad . \quad (2)$$

A similar expression describes the likelihood that subsequence $s^M[i \dots (i + \sigma(W_j) - 1)]$ of the mouse sequence is a binding site for the transcription factor j . The maximum over the positive and negative strand of the average of these two probabilities is the score S_i^j at position i for weight matrix j . Similarly, we take the score of position i under the background model B^H and B^M to be $S_i^B = (P(s^H[i] | B^H) + P(s^M[i] | B^M))/2$. A sequence alignment is parsed into binding sites and background. The likelihood of each parse is the product of the likelihoods of all binding sites and background regions in the parse. The sum of the likelihood over all parses is the partition function, $Z(1, L)$,

$$Z(1, L) = \sum_{\text{all parses}} \sum_{i=1}^L S_i^j,$$

where L is the length of the HSP, and the background is treated as a special weight matrix. The probability that we have a binding site of type j starting at position i in the HSP is

$$\frac{Z(1, i-1) S_i^{W_j} Z(i + \sigma(W_j), L)}{Z(1, L)}.$$

We then use a dynamic programming technique to compute the probability that the factor described by weight matrix W_j will bind at that position.

5 Acknowledgments

We thank Boris Lenhard for sharing data sets and results prior to publication, and Gabriela Loots for sharing sequence data. We are indebted to Arndt Benecke for a critical reading of the manuscript. This research has been supported in part by NINDS National Institutes of Health Grant NS39662, the Seaver Foundation and Albert Einstein College of Medicine (N. D. S), by National Cancer Institute Grant R33-CA84699 and National Science Foundation Grant DBI9984882 (T. G. and M. Z.), by The Rockefeller University Lita Annenberg Hazen Presidential Fellowship (M.Z.), and by NSF grant DMR 0129848 (N.R.).

References

- Abdollahi, A., Lord, K. A., Hoffman-Liebermann, B., and Liebermann, D. A. 1991. Sequence and expression of a cDNA encoding myd118: a novel myeloid differentiation primary response gene induced by multiple cytokines. *Oncogene* **6**: 165–167.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.
- Arnone, M. I., Bogarad, L. D., Collazo, A., Kirchhamer, C. V., Cameron, R. A., Rast, J. P., Gregorians, A., and Davidson, E. H. 1997. Green fluorescent protein in the sea urchin: new experimental approaches to transcriptional regulatory analysis in embryos and larvae. *Development* **124**: 4649–4659.
- Bailly, A., Torres-Padilla, M. E., Tinel, A. P., and Weiss, M. C. 2001. An enhancer element 6 kb upstream of the mouse hnf4alpha1 promoter is activated by glucocorticoids and liver-enriched transcription factors. *Nucl. Acids Res.* **29**: 3495–34505.
- Berg, O. G. and von Hippel, P. H. 1987. Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193**: 723–750.
- Bussemaker, H. J., Li, H., and Siggia, E. D. 2001. Regulatory element detection using correlation with expression. *Nat. Genet.* **27**: 167–171.

- Cassuto, H., Olswang, Y., Livoff, A. F., Nechushtan, H., Hanson, R. W., and Reshef, L. 1997. Involvement of hnf-1 in the regulation of phosphoenolpyruvate carboxykinase gene expression in the kidney. *FEBS Lett* **412**: 597–602.
- Chao, K. M. and Miller, W. 1995. Linear-space algorithms that build local alignments from fragments. *Algorithmica* **13**: 106–134.
- Chin, P. L., Momand, J., and Pfeifer, G. P. 1997. In vivo evidence for binding of p53 to consensus binding sites in the p21 and gadd45 genes in response to ionizing radiation. *Oncogene* **15**: 87–99.
- Coffer, P., Luttkicken, C., van Puijenbroek, A., Klop-de Jonge, M., Horn, F., and Kruijer, W. 1995. Transcriptional regulation of the junb promoter: analysis of stat-mediated signal transduction. *Oncogene* **10**: 985–994.
- Davuluri, R. V., Grosse, I., and Zhang, M. Q. 2001. Computational identification of promoters and first exons in the human genome. *Nat. Genet.* **29**: 412–417.
- De Smaele, E., Zazzeroni, F., Papa, S., Nguyen, D. U., Jin, R., Jones, J., Cong, R., and Franzoso, G. 2001. Induction of gadd45beta by nf-kappab downregulates pro-apoptotic jnk signalling. *Nature* **414**: 308–313.
- Dieterich, C., Cusack, B., Wang, H., Rateitschak, K., Krause, A., and M., V. 2002. Annotating regulatory DNA based on man–mouse genomic comparison. *Bioinformatics* **S2**: S84–S90.
- Gusfield, D. 1997. *Algorithms on strings, trees and sequences*. Press Syndicate of the University of Cambridge.
- Hardison, R., Oeltjen, J., and Miller, W. 1997. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.* **10**: 959–66.
- Hipskind, R. A., Baccharini, M., and Nordheim, A. 1994. Transient activation of raf-1, mek, and erk2 coincides kinetically with ternary complex factor phosphorylation and immediate-early gene promoter activity in vivo. *Mol Cell Biol* **14**: 6219–6231.
- Holmes, I. and Bruno, W. J. 2000. Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**: 202–210.
- Ishida, S., Huang, E., Zuzan, H., Spang, R., Leone, G., West, M., and Nevins, J. R. 2001. Role for e2f in control of both dna replication and mitotic functions as revealed from dna microarray analysis. *Mol Cell Biol* **21**: 4684–4699.
- Jareborg, N., Birney, E. ., and Durbin, R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Research* **9**: 815–824.
- Jegga, A., Sherwood, S., Carman, J., Pinski, A., Phillips, J., Pestian, J., and Aronow, B. 2002. Detection and visualization of compositionally similar cis–regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res.* **12**: 1408–1417.
- Jin, S., Fan, F., Fan, W., Zhao, H., Tong, T., Blanck, P., Alomo, I., Rajasekaran, B., and Zhan, Q. 2001. Transcription factors oct-1 and nf-ya regulate the p53-independent induction of the gadd45 following dna damage. *Oncogene* **20**: 2683–2690.
- Kastan, M. B., Zhan, Q., el Deiry, W. S., Carrier, F., Jacks, T., Walsh, W. V., Plunkett, B. S., Vogelstein, B., and Fornace, A. J. J. 1992. A mammalian cell cycle checkpoint pathway utilizing p53 and gadd45 is defective in ataxia-telangiectasia. *Cell* **71**: 587–597.
- Kitabayashi, I., Kawakami, Z., Matsuoka, T., Chiu, R., Gachelin, G., and Yokoyama, K. 1993. Two cis-regulatory elements that mediate different signaling pathways for serum-dependent activation of the junb gene. *J Biol Chem* **268**: 14482–14489.
- Kraus, W. L., Manning, E. T., and Kadonaga, J. T. 1999. Biochemical analysis of distinct activation functions in p300 that enhance transcription initiation with chromatin templates. *Mol Cell Biol* **19**: 8123–8135.

- Leone, G., DeGregori, J., Yan, Z., Jakoi, L., Ishida, S., Williams, R. S., and Nevins, J. R. 1998. E2f3 activity is regulated during the cell cycle and is required for the induction of s phase. *Genes Dev* **12**: 2120–2130.
- Loots, G., Locksley, R., Blankespoor, C., Wang, Z. E., Miller, W., et al. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**: 136–40.
- Loots, G., Ovcharenko, I., Pachter, L., Dubchak, I., and Rubin, E. 2002. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* **12**: 832–839.
- Matsumoto, A., Masuhara, M., Mitsui, K., Yokouchi, M., Ohtsubo, M., Misawa, H., Miyajima, A., and Yoshimura, A. 1997. Cis, a cytokine inducible sh2 protein, is a target of the jak-stat5 pathway and modulates stat5 activation. *Blood* **89**: 3148–3154.
- Mui, A. L., Wakao, H., Kinoshita, T., Kitamura, T., and Miyajima, A. 1996. Suppression of interleukin-3-induced gene expression by a c-terminal truncated stat5: role of stat5 in proliferation. *EMBO J* **15**: 2425–2433.
- O'Brien, R. M., Noisin, E. L., Suwanichkul, A., Yamasaki, T., Lucas, P. C., Wang, J. C., Powell, D. R., and Granner, D. K. 1995. Hepatic nuclear factor 3- and hormone-regulated expression of the phosphoenolpyruvate carboxykinase and insulin-like growth factor-binding protein 1 genes. *Mol Cell Biol* **15**: 1747–1758.
- Ohtani, K., Iwanaga, R., Nakamura, M., Ikeda, M., Yabuta, N., Tsuruga, H., and Nojima, H. 1999. Cell growth-regulated expression of mammalian mcm5 and mcm6 genes mediated by the transcription factor e2f. *Oncogene* **18**: 2299–2309.
- Pruitt, K. D. and Maglott, D. R. 2001. Refseq and locuslink: Ncbi gene-centered resources. *Nucleic Acids Research* **29**: 137–140.
- Ptashne, M. and Gann, A. 1999. Imposing specificity by localization: mechanism and evolvability. *Curr. Biol.* **8**: R812.
- Rajewsky, N., Vergassola, M., Gaul, U., and Siggia, E. D. 2002. Computational detection of *cis*-regulatory modules, applied to body patterning in the early *drosophila* embryo. *BMC Bioinformatics* **3**: 30.
- Rein, T., Muller, M., and Zorbas, H. 1994. In vivo footprinting of the irf-1 promoter: inducible occupation of a gas element next to a persistent structural alteration of the dna. *Nucleic Acids Res* **22**: 3033–3037.
- Riken Phase II Team and FANTOM Consortium. 2001. Functional annotation of a full-length mouse cDNA collection. *Nature* **409**: 685–690.
- Schwenzer, R., Siemienski, K., Liptay, S., Schubert, G., Peters, N., Scheurich, P., Schmid, R. M., and Wajant, H. 1999. The human tumor necrosis factor (tnf) receptor-associated factor 1 gene (traf1) is up-regulated by cytokines of the tnf ligand family and modulates tnf-induced activation of nf-kappab and c-jun n-terminal kinase. *J Biol Chem* **274**: 19368–19374.
- Sims, S. H., Cha, Y., Romine, M. F., Gao, P. Q., Gottlieb, K., and Deisseroth, A. B. 1993. A novel interferon-inducible domain: structural and functional analysis of the human interferon regulatory factor 1 gene promoter. *Mol Cell Biol* **13**: 690–702.
- Starr, R., Willson, T. A., Viney, E. M., Murray, L. J., Rayner, J. R., Jenkins, B. J., Gonda, T. J., Alexander, W. S., Metcalf, D., Nicola, N. A., and Hilton, D. J. 1997. A family of cytokine-inducible inhibitors of signalling. *Nature* **387**: 917–921.
- Suzuki, Y., Ishihara, D., Sasaki, M., Nakagawa, H., Hata, H., Tsunoda, T., Watanabe, M., Komatsu, T., Ota, T., Isogai, T., Suyama, A., and Sugano, S. 2000. Statistical analysis of the 5' untranslated region of human mRNA using oligo-Capped cDNA libraries. *Genomics* **64**: 286–297.
- Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A., and S., S. 1997. Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library. *Gene* **200**: 149–156.
- Takahashi, S., Saito, S., Ohtani, N., and Sakai, T. 2001. Involvement of the oct-1 regulatory element of the gadd45 promoter in the p53-independent response to ultraviolet irradiation. *Cancer Res* **61**: 1187–1195.

- Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y. 2001. A higher-order background model improves the detection of promoter regulatory elements by gibbs sampling. *Bioinformatics* **17**: 1113–1122.
- Thompson, B. J., Shang, C. A., and Waters, M. J. 2000. Identification of genes induced by growth hormone in rat liver using cDNA arrays. *Endocrinology* **141**: 4321–4324.
- Wang, C. Y., Mayo, M. W., Korneluk, R. G., Goeddel, D. V., and Baldwin, A. S. J. 1998. Nf-kappaB antiapoptosis: induction of traf1 and traf2 and c-iap1 and c-iap2 to suppress caspase-8 activation. *Science* **281**: 1680–1683.
- Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W., and Lawrence, C. E. 2000. Human-mouse genome comparisons to locate regulatory sites. *Nature Genetics* **26**: 225–8.
- Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R., Pruss, M., Schacherer, F., Thiele, S., and Urbach, S. 2001. The transfac system on gene expression regulation. *Nucleic Acids Res* **29**: 281–283.
- Xanthopoulos, K. G., Prezioso, V. R., Chen, W. S., Sladek, F. M., Cortese, R., and Darnell, J. E. J. 1991. The different tissue transcription patterns of genes for hnf-1, c/ebp, hnf-3, and hnf-4, protein factors that govern liver-specific transcription. *Proc. Natl. Acad. Sci. U. S. A.* **88**: 3807–3811.
- Yang, E., Henriksen, M. A., Schaefer, O., Zakharova, N., and Darnell, J. E. J. 2002. Dissociation time from DNA determines transcriptional function in a stat1 linker mutant. *J Biol Chem* **277**: 13455–13462.
- Yanuka-Kashles, O., Cohen, H., Trus, M., Aran, A., Benvenisty, N., and Reshef, L. 1994. Transcriptional regulation of the phosphoenolpyruvate carboxykinase gene by cooperation between hepatic nuclear factors. *Mol Cell Biol* **14**: 7124–7133.
- Zhan, Q., Carrier, F., and Fornace, A. J. J. 1993. Induction of cellular p53 activity by DNA-damaging agents and growth arrest. *Mol Cell Biol* **13**: 4242–4250.
- Zhao, R., Gish, K., Murphy, M., Yin, Y., Notterman, D., Hoffman, W. H., Tom, E., Mack, D. H., and Levine, A. J. 2000. Analysis of p53-regulated gene expression patterns using oligonucleotide arrays. *Genes Dev* **14**: 981–993.

6 Web references

- <http://hal.rockefeller.edu/Smash>. Annotation of putative transcription factor binding sites in the human genome using human-mouse homology.
- <http://www.gmod.org>. Generic Genome Browser developed by Lincoln Stein.
- <http://genome.ucsc.edu>. Human Genome Project.
- ftp://wolfram.wi.mit.edu/pub/mouse_contigs/MGSC_V3. Public assembly of the mouse genome.



Figure 1: Screenshot of SMASH annotation of the interferon regulatory factor locus.