BIOINFORMATICS

Vol. 17 no. 10 2001
Pages 871–877

Enrichment of regulatory signals in conserved non-coding genomic sequence

Samuel Levy [1,], Sridhar Hannenhalli [1] and Christopher Workman [2]*

[1]Informatics Research, Celera Genomics Corporation, 45 West Gude Drive, Rockville, MD 20850, USA and [2]Center for Biological Sequence Analysis, The Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark

Received on April 20, 2001; revised and accepted on July 6, 2001

ABSTRACT

Motivation: Whole genome shotgun sequencing strategies generate sequence data prior to the application of assembly methodologies that result in contiguous sequence. Sequence reads can be employed to indicate regions of conservation between closely related species for which only one genome has been assembled. Consequently, by using pairwise sequence alignments methods it is possible to identify novel, non-repetitive, conserved segments in non-coding sequence that exist between the assembled human genome and mouse whole genome shotgun sequencing fragments. Conserved non-coding regions identify potentially functional DNA that could be involved in transcriptional regulation.

Results: Local sequence alignment methods were applied employing mouse fragments and the assembled human genome. In addition, transcription factor binding sites were detected by aligning their corresponding positional weight matrices to the sequence regions. These methods were applied to a set of transcripts corresponding to 502 genes associated with a variety of different human diseases taken from the Online Mendelian Inheritance in Man database. Using statistical arguments we have shown that conserved non-coding segments contain an enrichment of transcription factor binding sites when compared to the sequence background in which the conserved segments are located. This enrichment of binding sites was not observed in coding sequence. Conserved non-coding segments are not extensively repeated in the genome and therefore their identification provides a rapid means of finding genes with related conserved regions, and consequently potentially related regulatory mechanism. Conserved segments in upstream regions are found to contain binding sites that are co-localized in a manner consistent with experimentally known transcription factor pairwise co-occurrences and afford the identification of novel co-occurring Transcription Factor (TF) pairs. This study provides a methodology and more evidence to suggest that conserved non-coding regions are biologically significant since they contain a statistical enrichment of regulatory signals and pairs of signals that enable the construction of regulatory models for human genes.

Contact: samuel.levy@celera.com

*To whom correspondence should be addressed.

INTRODUCTION

Regulatory regions of genes are notoriously difficult to identify in bulk genomic sequences using computational methods (Duret and Bucher, 1997; Fickett and Hatzigeorgiou, 1997; Bucher, 1999). These sequence regions, which for the most part reside in non-coding regions of genes, exhibit a regulatory function in transcription by virtue of containing binding sites for Transcription Factor (TF) protein or by chemical modifications on DNA that re-structure chromatin. Therefore a regulatory region could be characterized by the presence of a TF binding site(s) that typically vary from 5 to 20 bp in length, or by chemical moieties on DNA that are challenging to detect within bulk sequence. While TF binding sites are possible to characterize, their apparent lack of sequence complexity given their size, result in sequence search methods finding them on average every 1000 bp. This results in an unacceptably high false positive rate thereby reducing the value of TF binding site identification by these methods.

Another approach to this problem is to consider the conservation between relatively closely related species as a means to identify biologically significant DNA. Functionally relevant DNA sequences tend to accumulate mutations at a slower rate than neutral sequence through evolution. Sequence conservation is a measure of whether a sequence is biologically functional therefore a highly conserved sequence is more likely to carry with it an important biological role. These notions are well encapsulated in methods like ‘phylogenetic footprinting’, which permit

the identification of *cis*-acting regulatory elements by performing alignments of non-coding regions of orthologous genes and then identifying 100% sequence conservation in regions longer than six base pairs (bps) (Tagle *et al.*, 1988; Gumucio *et al.*, 1996). Comparative sequence analysis methods, using mouse and human genomic sequence, have identified regions of conserved non-coding sequence that can be shown experimentally to play a role in transcriptional control (Loots *et al.*, 2000; Wasserman *et al.*, 2000).

In order to identify a functional sequence for any species it is necessary to have a genomic sequence for at least one other species of sufficient evolutionary distance. In this manner it becomes important that the two or more species to be compared have accumulated sufficient mutation in neutral sequence but still be close enough in evolution to maintain common regulatory mechanisms (Duret and Bucher, 1997). Since obtaining sufficient sequence data used to pose more technical problems in the past, many approaches to date have applied a comparative genomics approach to a handful of genes on a locus for which sufficient sequence data was available (Hardison *et al.*, 1997; Oeltjen *et al.*, 1997). The rapid sequencing of higher eukaryotic genomes will permit genome-wide sequence comparision that can identify conserved non-coding sequences with regulatory capacity. In anticipation that the whole genome shotgun approach can produce many sequence reads, Bouck *et al.* (2000) proposed sequence comparison of these reads with a completely assembled genome as a means to identify functional sequence.

We describe a methodology that employs the assembled human genome and finds local alignments of upstream and intronic regions of identifiable human transcripts using mouse sequence reads. The resulting conserved segments identified on human sequence are subsequently analyzed for the presence of TF binding sites using known binding site sequences and positional weight matrices described from the biological literature and extracted from TRANSFAC (Wingender *et al.*, 2001). Using this methodology it was possible to automate the analysis of 502 genes taken from a disease gene database, Online Mendelian Inheritance in Man (OMIM). Using statistical arguments it was possible to show the enrichment of TF binding sites in conserved non-coding segments. Using stringent search methods to identify TF binding sites on genomic sequence it is also possible to infer potential interaction among TF proteins due to detectable co-localization of TF binding sites.

SYSTEM AND METHODS

Identifying conserved segments associated with human transcripts

We were able to map the mRNA for 502 uniquely determined OMIM genes to 566 transcripts identified by an extensive automated annotation process performed on the assembled human genome and described in more detail elsewhere (Venter *et al.*, 2001). The results of this preceding annotation step permitted us to identify three sequence classes from these transcripts, namely upstream (5 kbp relative to the translation start site), exon and intron. Using local sequence aligment methods (Altschul *et al.*, 1997) the alignment between mouse fragments (comprising $> 3\times$ mouse genome coverage) and the human OMIM gene sequence regions was determined, considering all alignments of 50 bp or longer at 70% identity or higher. By adjusting the BLAST alignment match and mismatch parameters, where a match $= +1$ and mismatch $= -1$ compared to the default values of $+1$ and -3 respectively, permitted the identification of aligned sequence blocks with as little as 50% identity. Improving alignment sensitivity is especially appropriate given that regulatory signals and/or the intervening sequence may not exhibit strong sequence conservation. All mouse fragments were masked to remove known rodent repeat sequence prior to performing alignments. After identifying non-redundant regions of conserved segments additional pairwise alignments were performed to remove potential neighbouring coding regions and to determine how repeated a particular conserved segment was in the human genome. The pipeline process to achieve novel conserved sequence regions in human genes is illustrated in Figure 1.

In general the process used for the identification of conserved segments between mouse and human avoids the inclusion of regions of known repeat sequence in both species (ALU, LINEs, etc.), since we are concerned with detecting novel sequence. This novel sequence is likely to contain the highest amount of gene-specific regulatory signals and is of primary concern. However, there is accumulating evidence to suggest that ALU repetitive elements play an as yet undetermined role in controlling gene expression (Santamarina-Fojo *et al.*, 2000; Willoughby *et al.*, 2000), indicating that future considerations of repeated elements between species may contribute to a further understanding of gene regulation.

Identifying transcription factor binding sites

Position Weight Matrices (PWMs) and binding site sequences, 10 bp or longer, representing TF binding sites known to occur in human were extracted from TRANSFAC version 4.4 (Wingender *et al.*, 2001). These data curation steps results in 170 PWMs and 1118 consensus sequence binding sites. Alignment of PWMs on sequences

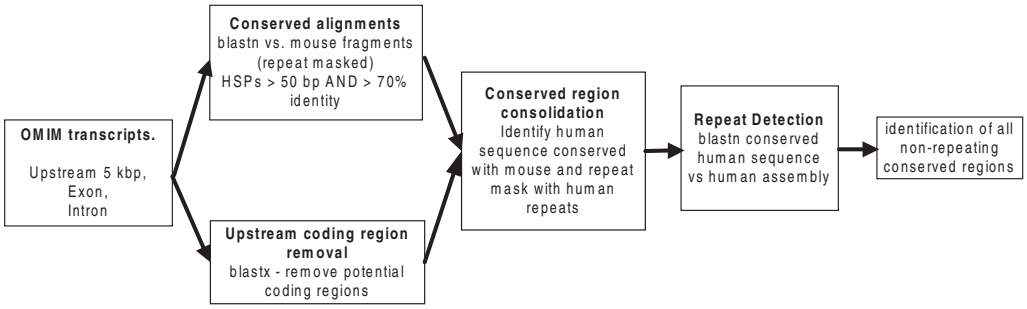

Fig. 1. Protocol for identifying novel conserved segments of human sequence using local alignment approaches and employing mouse sequence fragments.

was performed with PATSER (Hertz and Stormo, 1999) using *a priori* probability for sequence bases as calculated from the sequence being tested. PWM alignments with a probability of 10^{-6} or lower were accepted. Ungapped local sequence alignment methods, performed using BLAST, located binding site consensus sequences on a genomic sequence by ensuring an exact match of the binding site.

ALGORITHM

Z-score and *P*-value calculation

In the following we attempt to test the hypothesis of whether there is an enrichment of TF binding sites in conserved non-coding sequence which, if correct, provides a computational method to filter out 'noise' whilst searching for TF binding sites. A significance measure was associated with the frequency of occurrence of bps in TF binding sites within the conserved genomic sequence as compared to all genomic sequences of a certain class (e.g. upstream, exon and intron), as a background.

For example, consider we are analyzing the upstream region. If c is the fraction of the upstream region that is conserved, then among all the TF binding site bp within the upstream region, we would expect the c fraction to fall within the conserved region, if there was no correlation between conservation and TF binding. Using a simple binomial model with n events, where n is the total number of bases accounting for all the TF binding site bp in the upstream region, and using c as the probability of success, we can calculate the expected number of successes, mean $u = n \times c$, and the standard deviation $\sigma = \sqrt{n \times c \times (1-c)}$. Let t be the fraction of TF binding site bp within the conserved region and from this the Z-score is calculated as

$$Z = \frac{t-u}{\sigma}.$$

Then the probability of observing t successes, as given by Chebyshev's Theorem is

$$\left(\frac{t-u}{\sigma}\right)^{-2}.$$

For each sequence class and for both TF PWMs and experimentally determined TF binding sites, we calculated this probability (or P-value). The lower the P-value, the greater the evidence that conserved regions have an enrichment of TF sites.

Transcription factor co-localization

Let n_i be the number of hits of a TRANSFAC PWM corresponding to a particular factor i, in a set of sequences S. For a pair of TFs i and j, let N_{ij} be the number of times, the two PWM hits occur within w bps in S. In order to provide a significance of the measured co-occurrence of factors, an appropriate randomly generated background co-occurrence model is necessary. There are several potential background models that can be generated. We shuffled randomly the n_i number of hit positions for factor i and n_j number of hit positions for factor j in S. Then let R_{ij} be the number of times, the two PWMs hit within w bps in S, using the randomly shuffled TF site locations. This background co-occurrence model therefore generates random positions of the observed number of pooled TF binding sites across the region of sequence tested. Another potential background model would be to only generate random positions for the observed type and number of binding sites for any one sequence and to perform this randomization within each sequence region in our data set. Once the foreground and background co-localization events have been counted the preferential co-localization of a pair of factors i and j is measured in terms of Co-localization Index, CI, defined as:

$$\mathrm{CI} = \log \frac{N_{ij}}{R_{ij}}.$$

Notice that CI is symmetric for a pair of factors, resulting in $(n^2-n)/2$ values, where n is the total number of PWMs

from TRANSFAC. Values of $w = 50$, 100 and 200 bp were employed for the window size and sites occurring on both strands were considered equally. In addition, pairs of PWMs corresponding to pairs of TFs known to co-localize based on experimental evidence from the biological literature, were extracted from TRANSFAC version 4.4. This resulted in 261 pairs of PWMs, involving 77 individual PWMs, and constituted a literature-curated set of pairs for which CI could be measured and compared with CI for all PWM pairs. CI values were obtained for TF pairs identified in conserved segments only and also in whole upstream regions disregarding conservation. There are instances where two factors have similar sequence preferences, as evidenced by possessing similar matrices or consensus sequences. In these situations we would expect artifactual false positive CI scores. However these similar factors constitute a small number in TRANSFAC and given the large number of TF pairs tested, these false positive pairs would constitute a small percentage of all pairs calculated.

RESULTS

Conservation degree

The fraction of sequence regions that is conserved are shown in Table 1. The highest conservation occurs in exonic sequence for the OMIM genes (79%) but only 20 and 24% of the 500 bp and 5 kbp upstream of the first coding exon is conserved respectively. Introns and specifically the first intron, where some regulatory signals have been shown to reside (Storbeck *et al.*, 1998), exhibit the least degree of sequence conservation. However it should be noted that in terms of raw conserved sequence, introns contain almost twice as much sequence as upstream. The average length of conserved segments in upstream and intron regions is approximately 150 bp while conserved exon segments are around 180 bp.

The degree of conservation in each sequence class does vary as indicated by Figure 2. Exon conservation is evenly distributed from 70 to 95% identity with a peak of conservation around 80%. Non-coding sequence conservation is spread between 70 and 85% identity with a peak of 70% for intron and 75% for upstream. This is consistent with the notion that coding sequence results in highly conserved amino acid sequences that is detectable between mouse and human. Conversely, regulatory signals are conserved at the DNA level where a lesser degree of sequence conservation would not interfere with the binding of the same regulatory proteins that control gene expression in orthologues.

Transcription factor binding sites in conserved segments

We attempted to answer the question whether the conserved non-coding segments are more likely to contain TF binding sequence. To this end, we determined the exact match of sequences 10 bp or greater in length that have been characterized in TRANSFAC (Wingender *et al.*, 2001) as TF binding sites, to the sequence regions of the OMIM genes. Upon alignment it is possible to compute the percentage of the TF binding sequence that occurs within conserved regions (see Table 1, column 3) in a particular sequence class and compare this to the percentage conserved sequence within that sequence class (Table 1, column 2) since this is the probability of seeing a site aligned by random chance within the conserved segments of the sequence class. The Z-score indicates whether TF binding sites are over or underrepresented in conserved regions versus the distribution of binding sites in the whole sequence region. From this comparision it is possible to see that all conserved segments in non-coding regions have an over-representation of TF binding sites (large Z-score). A similar trend is seen when aligning TRANSFAC PWMs, rather than the TF sequences (Table 1, columns 5 and 6). Alternatively, coding regions

Table 1. Conservation of upstream (500 bp and 5 kbp), exon, intron and the first intron from 502 OMIM genes

		TF sequences aligned to conserved segments		TF matrices aligned to conserved segments	
	Conserved bp (%)[1]	(%)[2]	Z-score, P-value[3]	(%)[4]	Z-score, P-value
Upstream (500 bp)	20	45	23.9, 0.0018	26	9.3, 0.0115
Upstream (5 kbp)	24	43	32.0, 0.0009	28	10.9, 0.0084
Exon	79	73	−5.0, 0.0407	77	−4.7, 0.0449
Intron—all	12	18	25.6, 0.0015	17	62.6, 0.0003
Intron—first	10	15	14.1, 0.0051	16	42.4, 0.0006

[1]Conserved base pairs found in each sequence class obtained by aligning mouse fragments (representing $> 3\times$ coverage of mouse genome) expressed as percentage of sequence class (i.e. number of conserved bp in class/number all bp in class *100%).
[2]Percentage of experimentally determined TF binding sites taken from TRANSFAC version 4.4 that occur in conserved segments (i.e. number of TF site bp occurring in conserved segments/number of TF site bp in class *100%). Experimental sites were 10 bp or longer and were aligned to the entire sequence in each class.
[3]Z-score and P-value reflect extent and confidence of over-representation of TF binding site positions in conserved segments.
[4] Percentage of bases aligned to PWMs taken from TRANSFAC (constructed with human data) that occur within conserved segments using high stringency (P-value of matrix alignment $< 1 \times 10^{-6}$). (Calculated in the same manner as TF sequences aligned to conserved segments.)

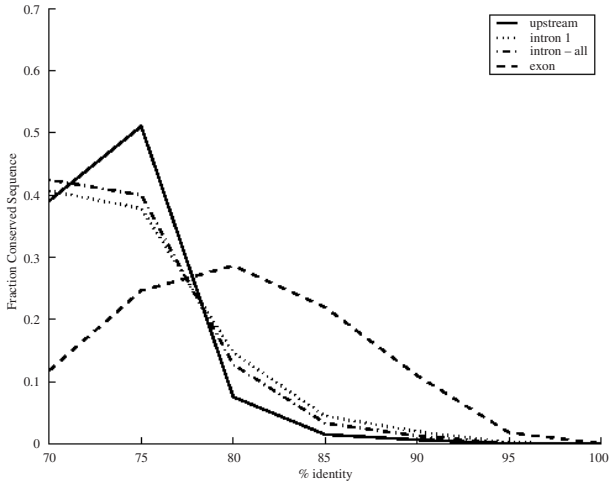

Fig. 2. Sequence conservation found in conserved segments identified by alignments of mouse fragments with human genomic sequence indicates that conserved exonic sequence displays a higher degree of conservation than conserved non-coding sequence i.e. upstream and intronic regions.

have a slight underrepresentation of TF binding sites suggesting that there is no enrichment of binding sites within conserved exon segments. It is interesting to note that a significant over-representation of TF binding sites occur in conserved segments from introns as in upstream regions. Taken together, these results suggest that the conserved non-coding sequence contain an enrichment of sequence elements that are consistent with the presence of regulatory signals. In addition to known signals it is also likely that these sequence segments contain yet to be determined regulatory signals. Thus by reducing our effective search space to only consider conserved non-coding sequence, one can postulate a better likelihood of finding these concealed signals.

Repeated occurrence of conserved segments

In an attempt to characterize the uniqueness of the conserved segments, their occurrence was determined in the whole genome by local sequence alignments. Any conserved segment that contained either known human repeat motifs for more than 50% of its length, or less than 50 bp of non-repeat sequence, was eliminated from further consideration. This removal of conserved segments with known repeats reduces the number of conserved bp in 5 kbp upstream and intron regions by half since the percentage of conserved bp goes from 24 to 12% for upstream 5 kbp and from 12 to 7% for introns (compare columns 2 of Table 1 with 2). The number of occurrences of the remaining non-repeat motif containing conserved segments in the entire human genome was determined. The resulting analysis therefore describes the extent of low level repeat occurrence of conserved segments for each sequence class.

Table 2. Repeat level of conserved segments

	Conserved bp (%) after known repeat removal	Unique bp (%)	Repeated 2 × (%)	Repeated 3 × −10× (%)
Upstream (5 kbp)	12	56	22	15
Exon	73	44	26	26
Intron—all	7	59	24	14
Intron—first	7	58	24	11

The results of this analysis (Table 2) indicated that between 56 and 59% of all non-coding conserved segment bps were unique in the genome and an additional 22–24% were found at only two locations in the genome. All sequence classes contain conserved segments that occur between 3 and 10 times in the genome. In the context of non-coding sequence, this raises an interesting possibility that these other conserved segments, could contain functional regulatory sequence that are common for a number of genes. Clearly further characterization of the location of the repeats and their proximity to genes will enable a genome wide identification of genes with common regulatory elements.

Co-localization of transcription factors

TFs are likely to interact to form complexes that regulate transcription. Therefore we attempted to describe a possible co-localization event by calculating a CI, for each TF pair. Figure 3 shows the distribution of CI scores for all TF pairs found in the whole 5 kbp upstream region reported as ‘All TF pairs’ using a window size of 100 bp. The distribution is clearly skewed towards higher log-likelihood scores suggesting that in general there are more preferred co-localization events than one would expect by random chance. This is consistent with the notion that TFs bind in clusters forming transcription modules. Some TF pairs are known to co-localize by virtue of experimental determination and are reported as such in TRANSFAC. The CIs that correspond to these factor pairs were identified and reported in Figure 3 as ‘Interacting TF pairs’. This distribution is bi-modal with one peak shifted toward high log-likelihood values vindicating that high CI scores are a good indicator for interacting TF pairs. The same trend for CI was found when using just the TF pairs found in conserved segments and when window sizes of 50 and 200 bp were employed. Only 10% of the pairs of PWMs with known interactions have a CI $\geqslant 2$. Using this value of CI as the threshold, we could identify 214 novel PWM-

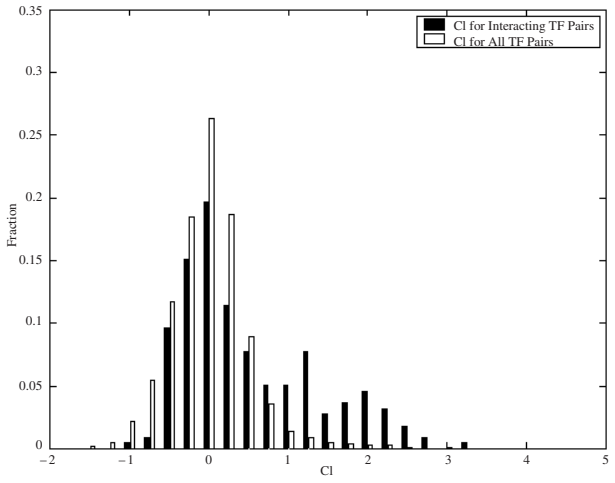

Fig. 3. Distribution of CI, computed from PWMs aligned to upstream 5 kbp regions of OMIM genes, using a window size of 100 bp. 'Interacting TF Pairs' is a subset of experimentally determined co-localized factors extracted from TRANSFAC identified amongst 'All TF pairs' vindicating that high scoring TF pairs correspond with known interacting factors from the literature.

pairs corresponding to 90 PWMs that may constitute novel interacting TF pairs that could form homo (91 pairs)—and hetero-meric (123 pairs) complexes. Clearly, this requires further investigation and corroboration from the published literature and additional experimental study.

DISCUSSION

We have attempted to show that by using the genomic sequence from two evolutionarily comparable higher eukaryotes it is possible to identify conserved regions of sequence in non-coding regions proximal to identifiable transcripts. Furthermore, these conserved segments can be shown to possess an enrichment of TF binding sites by alignment of known TF signal sequences. Whilst regulatory signals have been shown to occur upstream of genes, our analysis has shown a rather significant over-representation of TF binding sites in the conserved segments found in introns, and not just the first intron as is commonly expected. If this finding is corroborated experimentally then this indicates that intronic sequence signals could additionally control gene expression as a post-transcriptional event (Virt and Raschke, 2001).

In this analysis the transcripts and associated non-coding sequence were taken from the human genome and the mouse sequence was in the form of whole genome shotgun sequence reads. This approach suggests a potential future use of sequence reads prior to assembly (Bouck *et al.*, 2000) and the potential utility in employing sequence reads from low-pass sequence coverage of any particular genome. In order to ensure that using mouse sequence reads would provide sufficient sequence context and that no sequence conservation was missed, a preliminary study was performed to identify a set of muscle specific promoter regions for which known TF binding site data exists (Wasserman and Fickett, 1998; Wasserman *et al.*, 2000). We were able to locate all known muscle specific TF binding sites reported in the Wasserman *et al.* study in conserved segments using the BLAST-based search methodology outlined in Figure 1 (data not shown).

Previous studies have suggested that 'phylogenetic footprinting' would be effective in finding regulatory signals on the basis of analyzing orthologues in closely related species (Oeltjen *et al.*, 1997; Hardison *et al.*, 1997; Dubchak *et al.*, 2000). We have been able to apply a modified version of this technique in an automated way to a larger set of genes taken from the OMIM database that had been mapped to the human genome assembly sequence (Venter *et al.*, 2001). The initial goal was to identify gene specific regulatory signals and therefore we chose to eliminate known repeat sequence in mouse but only eliminated human repeats as a post-processing step once the conserved segments themselves were identified. Recent evidence suggest that some repeat motifs may themselves play a role in transcriptional regulation (Hamdi *et al.*, 2000), therefore their identification in conserved segments was deemed worthwhile. After eliminating known human repeat motifs from the conserved non-coding segments our analysis shows that 56–59% of conserved segment bp are unique in the genome. This suggests that some genes could possess a unique spatial distribution of TF binding sites if, in fact, the conserved segments contain relevant sites. This finding may have a therapeutic application by indicating specific ways in which to target the regulation of these genes. Of more immediate interest are the conserved segments that are repeated from 2 to 10 times in the genome (Table 2). This could indicate that several genes share a similar set of regulatory switches and therefore suggests a testable hypothesis that these genes could be co-expressed under certain conditions.

An important aspect of understanding how TF proteins interact and the potential role they play in forming transcriptional active units is to be able to identify all the interacting proteins. If certain proteins have a tendency to interact then one may expect to find their binding sites co-occurring in genomic DNA in an over-represented manner. This was the motivation for establishing a CI in order to compute the significance of observing the co-occurrence of TRANSFAC PWMs. Furthermore since conserved non-coding segments may not be the exclusive location of TF binding sites, TF site co-localization events were determined across whole upstream regions. Using

this approach it was possible to identify 214 novel putative interactions.

Finally, while our analysis supports the hypothesis that conserved non-coding segments contain an enrichment of TF binding sites, data in the literature show that conserved non-coding regions may not contain binding sites but still play a role in gene regulation. For example, a recent study on the sequence conservation in the upstream region of IL-4 and IL-13 genes indicate that DNase I hypersensitive sites can be found in conserved segments (Loots *et al.*, 2000). This result suggests that in this particular case sequence conservation is being employed to promote the absence of DNA–protein interactions, hence DNase I hypersensitivity. The ability to perform high stringency PWM alignments for known TF binding sites in all conserved non-coding segments or the application of motif finding techniques (Hertz and Stormo, 1999; Lawrence *et al.*, 1993; Bailey and Elkan, 1995) will indicate the presence or absence of regulatory signals in these regions. The application of these methods to only 10–20% of genomic sequence will appreciably reduce the false positive rate frequently associated with detecting regulatory regions.

ACKNOWLEDGEMENTS

S.L. and S.H. would like to thank Gene Myers, Vineet Bafna of Celera Genomics, USA and Gary Stormo, Washington University, USA, for their many valuable comments during the course of this study. C.T.W. is funded by the Danish National Research Foundation.

REFERENCES

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.

Bailey,T.L. and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. *Ismb*, **3**, 21–29.

Bouck,J.B., Metzker,M.L. and Gibbs,R.A. (2000) Shotgun sample sequence comparisons between mouse and human genomes. *Nature Genet.*, **25**, 31–33.

Bucher,P. (1999) Regulatory elements and expression profiles. *Curr. Opin. Struct. Biol.*, **9**, 400–407.

Dubchak,I., Brudno,M., Loots,G.G., Pachter,L., Mayor,C., Rubin,E.M. and Frazer,K.A. (2000) Active conservation of non-coding sequences revealed by three-way species comparisons. *Genome Res.*, **10**, 1304–1306.

Duret,L. and Bucher,P. (1997) Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.*, **7**, 399–406.

Fickett,J.W. and Hatzigeorgiou,A.G. (1997) Eukaryotic promoter recognition. *Genome Res.*, **7**, 861–878.

Gumucio,D.L., Shelton,D.A., Zhu,W., Millinoff,D., Gray,T., Bock,J.H., Slightom,J.L. and Goodman,M. (1996) Evolutionary strategies for the elucidation of *cis* and *trans* factors that regulate the developmental switching programs of the beta-like globin genes. *Mol. Phylogenet. Evol.*, **5**, 18–32.

Hamdi,H.K., Nishio,H., Tavis,J., Zielinski,R. and Dugaiczyk,A. (2000) Alu-mediated phylogenetic novelties in gene regulation and development. *J. Mol. Biol.*, **299**, 931–939.

Hardison,R.C., Oeltjen,J. and Miller,W. (1997) Long human–mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.*, **7**, 959–966.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Loots,G.G., Locksley,R.M., Blankespoor,C.M., Wang,Z.E., Miller,W., Rubin,E.M. and Frazer,K.A. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons [see comments]. *Science*, **288**, 136–140.

Oeltjen,J.C., Malley,T.M., Muzny,D.M., Miller,W., Gibbs,R.A. and Belmont,J.W. (1997) Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res.*, **7**, 315–329.

Santamarina-Fojo,S., Peterson,K., Knapper,C., Qiu,Y., Freeman,L., Cheng,J.F., Osorio,J., Remaley,A., Yang,X.P., Haudenschild,C., Prades,C., Chimini,G., Blackmon,E., Francois,T., Duverger,N., Rubin,E.M., Rosier,M., Denefle,P., Fredrickson,D.S. and Brewer, Jr,H.B. (2000) Complete genomic sequence of the human ABCA1 gene: analysis of the human and mouse ATP-binding cassette a promoter. *Proc. Natl Acad. Sci. USA*, **97**, 7987–7992.

Storbeck,C.J., Sabourin,L.A., Waring,J.D. and Korneluk,R.G. (1998) Definition of regulatory sequence elements in the promoter region and the first intron of the myotonic dystrophy protein kinase gene. *J. Biol. Chem.*, **273**, 9139–9147.

Tagle,D.A., Koop,B.F., Goodman,M., Slightom,J.L., Hess,D.L. and Jones,R.T. (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.*, **203**, 439–455.

Venter,J.C. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Virt,E.L. and Raschke,W.C. (2001) The role of intronic sequence in high level expression from CD45 cDNA constructs. *J. Biol. Chem.*, **276**, 19 913–19 920.

Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.

Wasserman,W.W., Palumbo,M., Thompson,W., Fickett,J.W. and Lawrence,C.E. (2000) Human–mouse genome comparisons to locate regulatory sites. *Nature Genet.*, **26**, 225–228.

Willoughby,D.A., Vilalta,A. and Oshima,R.G. (2000) An Alu element from the K18 gene confers position-independent expression in transgenic mice. *J. Biol. Chem.*, **275**, 759–768.

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.