

Protein-based analysis of alternative splicing in the human genome

Ann E. Loraine, Gregg A. Helt, Melissa S. Cline, and Michael A. Siani-Rose

Affymetrix, Inc., 6550 Vallejo Street, Emeryville, CA 94608 USA

ann_lorraine@affymetrix.com, gregg_helt@affymetrix.com, michael_siani-rose@affymetrix.com

Abstract

Understanding the functional significance of alternative splicing and other mechanisms that generate RNA transcript diversity is an important challenge facing modern-day molecular biology. Using homology-based, protein sequence analysis methods, it should be possible to investigate how transcript diversity impacts protein structure and function. To test this, a data mining technique ("DiffHit") was developed to identify and catalog genes producing protein isoforms which exhibit distinct profiles of conserved protein motifs. We found that out of a test set of over 1,300 alternatively spliced genes with solved genomic structure, over 30% exhibited a differential profile of conserved InterPro and/or Blocks protein motifs across distinct isoforms. These results suggest that motif databases such as Blocks and InterPro are potentially useful tools for investigating how alternative transcript structure affects gene function.

1. Introduction

Understanding how alternative splicing in the human genome affects protein structure and function is an important challenge facing modern-day molecular biology. Current estimates project that around half of all multi-exon, human genes give rise to more than one transcript variant [1], but the functional significance of this variation is unknown in most cases. Alternative splicing has been studied intensively at the level of individual genes, and examples of cell type and stage-specific expression of alternative mRNA species produced by the same loci are common in the scientific literature [2, 3]. Clearly, the pattern of splicing for many genes is highly regulated and differs from cell type to cell type.

Other mechanisms besides alternative splicing also contribute to transcript diversity. For example, alternative promoter choice, resulting in differential transcription start site selection, can produce transcripts with variable 5' regions. Likewise, alternative polyadenylation site choice can produce RNA species with variable 3' regions. These mechanisms can affect gene function either at the level of RNA, such as by differential inclusion of sequences controlling message localization or stability, or at the level of the translated protein product when coding regions are affected. Current estimates report that transcript variation

affects the coding region roughly three-fourths of the time [4]. Here, we explore the effects of those changes using the InterPro and Blocks profile libraries to annotate and compare distinct isoforms produced by the same gene.

Alternative transcript structure can affect protein function in at least two ways. First, functionally important coding sequence can be added or deleted, resulting in proteins with different, even antagonistic functions. For example, the BclX gene produces at least two distinct variants, one which inhibits apoptosis and another which promotes it [5]. Second, alternative transcript structure could simply remodel domains present in all isoforms, such as by mutually exclusive use of same-frame cassette exons, each encoding variations on the same general motif or functional domain. This is the case for Drosophila gene DSCAM, an immunoglobulin superfamily member involved in neuronal development [6].

Numerous libraries and methods for classifying proteins according to their homologies to known amino acid patterns or motif signatures have been developed. For example, the Blocks library of protein family profiles is based on conserved, ungapped segments ("Blocks") that typically occur in groups within related proteins [7]. The InterPro database organizes protein sequence profiles from several different protein analysis databases into a single resource and so provides possibly the most comprehensive description of the known protein universe currently available for public use [8]. Profiles which detect similar patterns have been grouped into single InterPro entries, and many of these InterPro "meta-motifs" have been assigned biological function using the Gene Ontology consortium's controlled vocabulary [9]. Using the program InterproScan, one can easily search for homologies to protein motifs described by the PRINTS, ProDom, PROSITE, Pfam and other InterPro member databases.

These and other protein classification systems have been widely used to predict function and family affiliation for newly discovered proteins deduced from the conceptual translation of genomic DNA and cDNA sequences. Given the sensitivity and wide scope of these methods, it is possible that they could also be used to study how alternative splicing impacts protein structure and function. By comparing the pattern of functional motifs detected in different protein isoforms produced the same gene, it should be possible to construct

hypotheses describing how alternative transcript structure may affect gene function.

To determine whether this approach is feasible, we used Blocks and InterPro member libraries to annotate a collection of proteins arising from genes with solved genomic structure that express two or more distinct transcript variants. In addition to providing a genome-level analysis of the impact of alternative transcript structure on protein structure and function, our results also summarize the current state of homology-based, protein sequence analysis methods and offer a first look at how conserved motifs are affected by alternative transcript structure.

2. Results

2.1. Genomic analysis

A collection of human cDNA sequences was downloaded from the RefSeq and GenBank databases available at NCBI. The downloaded sequences included 10,422 sequences from the RefSeq database [10] and 16,015 sequences from the larger GenBank database that were annotated as having a complete CDS and therefore were believed by their submitters to encode a full-length protein. The 26,437 cDNA sequences were aligned to working draft human genomic sequence (April, 2001 release) using pslayout, a cDNA-to-genomic sequence alignment program [11]. Using CDS annotations to delimit coding regions, a conceptual translation of genomic sequence was attempted for each aligned cDNA. The resulting collection of genome-derived protein translations included 18,190 sequences.

Because of the draft nature of the human genome sequence, it was then necessary to identify which of these proteins originated from genes located in genomic regions for which the sequence data is of relatively high quality. To identify these proteins, the pairwise alignment program bl2seq was used to align each conceptual translation against the corresponding protein sequence from GenPept.

All sequences that failed to align with 95% identity or better across the full length of both alignment partners were discarded, leaving a total of 17,811 sequences with a reliable genomic structure in the protein-coding region of each transcript.

2.2. Gene and transcript variant classification

The 17,811 high-quality transcript alignments were then grouped into genes and transcript groups by comparing their genomic alignments. For the purposes of this analysis, a gene is defined as a collection of transcripts in which all members have some protein-coding sequence in common and so encode isoforms of

the same protein. Here, transcripts were collected into gene groups according to whether they share at least 50 bases of in-frame, continuous coding sequence. Thus, all transcripts from the same gene group are related in the sense that they share at least 15 amino acids in common. Using this criterion, the 17,811 transcripts were assembled into 8,317 genes. In 187 cases, the same transcript was assigned to more than one gene. These genes were discarded, leaving 8,223 genes.

Next, transcripts that aligned to genomic sequence in a similar fashion were assigned to the same transcript structure group. That is, if two transcripts exhibited an identical pattern of alignment to genomic sequence across all inferred, internal splice boundaries, then they were assigned to the same transcript structure group and were treated as the same variant in subsequent analyses. Conversely, if an inferred intron in one transcript alignment overlapped an inferred exon in another, then the two transcripts were assigned to different transcript structure groups. This permitted proper group assignment for transcripts with incomplete sequence data at their 5' or 3' ends even when the original submitters believed that the full protein coding sequence was present.

To build a collection of genes with multiple alternative forms, genes containing just one transcript structure group were purged. The resulting collection contained 1,378 genes known to produce two or more distinct transcript variants as determined by their relative alignments to genomic sequence.

2.3. Protein sequence analysis

We next applied Blocks and InterPro to investigate how alternative transcript structure impacts protein structure and function. Specifically, we asked how often alternative transcript structure changed the pattern of motifs in the different protein isoforms belonging to the same gene.

To answer this question, a method (DiffHit) was developed to find cases where differences between alternative transcripts cause biologically meaningful differences in their encoded proteins. The inputs to this method are (1) a database of protein annotations, such as regions of conserved sequence detected by Blocks or InterPro, and (2) a database of gene and transcript structure group assignments for the transcripts and genes to be considered.

This method compares protein annotations associated with transcripts from different transcript structure groups belonging to the same gene and reports on any differences that are found. That is, it finds cases where variations in transcript structure coincide with variations in protein annotations and reports these.

Table 1. InterPro homologies ranked by number of genes matched.

(a) **InterPro homologies found in 6,017 (out of a possible 8,223) genes with reasonably well-solved genomic structure.** The full set, which included genes expressing just one as well as genes expressing multiple isoforms, matched 1,504 different InterPro profiles.

	Interpro Entry	# genes
1	Proline-rich region: <i>IPR000694</i>	457
2	Zinc finger, C2H2 type: <i>IPR000822</i>	239
3	Rhodopsin-like GPCR superfamily: <i>IPR000276</i>	198
4	Eukaryotic protein kinase: <i>IPR000719</i>	188
5	Serine/Threonine protein kinase: <i>IPR002290</i>	186
6	Tyrosine protein kinase: <i>IPR001245</i>	176
7	Immunoglobulin and major histocompatibility complex domain: <i>IPR003006</i>	166
8	Immunoglobulin subtype: <i>IPR003599</i>	123
9	Homeobox domain: <i>IPR001356</i>	107
10	G-protein beta WD-40 repeats: <i>IPR001680</i>	101

(b) **Most frequently occurring InterPro matches affected by alternative transcript structure.** The number of genes in which the indicated InterPro profile was affected by alternative transcript structure is shown in column 3.

	Interpro Entry	# alt diffs
1	Proline-rich region: <i>IPR000694</i>	27
2	Immunoglobulin and major histocompatibility complex domain: <i>IPR003006</i>	19
3	G-protein beta WD-40 repeats: <i>IPR001680</i>	14
4	Immunoglobulin subtype: <i>IPR003599</i>	14
5	Eukaryotic protein kinase: <i>IPR000719</i>	13
6	Zinc finger, C2H2 type: <i>IPR000822</i>	12
7	Immunoglobulin C-2 type: <i>IPR003598</i>	8
8	Serine/Threonine protein kinase: <i>IPR002290</i>	8
9	Tyrosine protein kinase: <i>IPR001245</i>	7
10	RNA-binding region RNP-1 (RNA recognition motif): <i>IPR000504</i>	5

First, the method finds cases where the type of motifs detected vary across variants. For example, if one variant lacks a motif, such as a membrane-spanning region, that is present in another, then the DiffHit method would report this. The method also finds cases where variants contain the same motif type, but disagree on the number of motif segments or spans. For example, a protein variant containing a different number of repeated, Zn-finger motifs than are found in alternative isoforms from the same gene would be reported. However, this method does not detect cases where motifs that are typically present as single, continuous spans of amino acid sequence and the recognized motif is merely longer in some isoforms but not in others.

The protein isoforms in the full test set of 8,223 genes were searched against the InterPro and Blocks databases using default parameter settings. The Blocks analysis detected conserved motifs in 5,389 (66%) of these genes, while the InterPro method recognized 6,017 (73%). Next, the subset of 1,378 genes containing two or more distinct transcript groups was examined. Genes where the Blocks or InterPro hits varied among the different transcript groups were noted. In the case of Blocks, 340 genes contained transcripts with differing patterns of motifs. The InterPro analysis identified 366 genes where the motifs differed among transcript groups. Together, both methods recognized a differing pattern of motifs in 475 genes. Thus alternative transcript structure was associated with detectable changes in motif structure for approximately 34% of the 1,378 genes containing two or more distinct transcript variants. The remainder either encoded a single protein isoform because alternative transcript structure affected the non-coding regions of the gene or because conserved regions in the encoded proteins were unaffected.

Next, we assessed whether some motifs are more frequently affected by alternative transcript structure than others. We compiled a list of the motifs observed most often when variations in transcript structure led to variations in the pattern of motifs, and compared this to the list of InterPro entries observed most often overall. The results were very similar for Blocks and InterPro; the InterPro results are shown here.

Table 1(a) lists the number of genes recognized by the top ten most frequently occurring InterPro profiles in the test set of 8,223 genes, in which 6,845 genes produced only one variant and 1,378 genes produced two or more variants. Table 1(b) lists the number of genes in which the indicated InterPro profile was affected by alternative transcript structure. In general, the two lists presented in Table 1 are very similar. Since alternative splicing is thought to affect about half of all human genes, it is reasonable to expect that the distribution of profiles affected by alternative transcript structure would resemble

that for all genes as whole. Our results support this expectation.

2.4. Gene-level analysis

Gene-level analysis was then done for the 475 genes in which different isoforms had different patterns of motif hits. To facilitate by-hand analysis, an interactive, Java-based visualization tool (ProtAnnot) was developed which displays protein motifs superimposed on an alignment between alternative transcripts and genomic sequence. As with other genome browser applications, exons are shown as rectangles connected by lines indicating introns. Translated regions are shown as filled rectangles, while untranslated 3' and 5' regions are shown as empty rectangles. Unlike other such browsers, ProtAnnot shades coding regions to indicate the frame of translation. ProtAnnot also summarizes the number of transcripts which contain exonic sequence at each base position as a series of blocks of varying height shown at the bottom of the display. These features help speed up visual inspection of each locus, allowing one to assess at a glance how intron-exon structure affects the encoded proteins. Using this tool, hypotheses regarding the function of alternative transcript structure were developed for some of the genes where the functional significance of the affected motif is known, and two examples of these are presented below.

2.4.1. MEOX1. MEOX1 (mesenchyme homeobox 1, also called MOX1) is located in the BRCA1 region on chromosome 17 and is the human homolog of the mouse Mox1 homeobox-containing gene [12]. MEOX1 is known to be involved in axial skeleton development and is reported to interact with the homeobox-containing transcription factor Pax1 [13]. MEOX1 encodes at least two distinct protein isoforms, a longer form (Figure 1, transcript (b)) recognized by Blocks and InterPro homeobox domain profiles and a shorter form (Figure 1, transcript (a)) that contains no homologies of any type. Both isoforms share a common 157 amino acid N-terminal region but differ in their C-termini.

Homeobox-containing proteins bind to DNA via the homeodomain often as hetero- or homodimers in complex with other homeodomain-containing proteins [14]. Thus the longer MEOX1 isoform appears competent to interact with DNA and therefore is probably capable of regulating transcription, while the shorter isoform most likely is not since it lacks a homeodomain. However, if the N-terminal region of MEOX1 permits the protein to interact with other proteins involved in regulating MEOX1 function, such as Pax1, then the shorter form which lacks a homeobox but which retains the amino-terminal region may serve as a negative regulator of MEOX1 function.

Alternatively, the short form of MEOX1 may represent an aberrant splicing event and may have no function. To distinguish these possibilities, further sequencing and functional analysis of MEOX1 cDNAs should be done to determine whether the shorter form is a true variant and whether it is competent to interact with known MEOX1 protein partners.

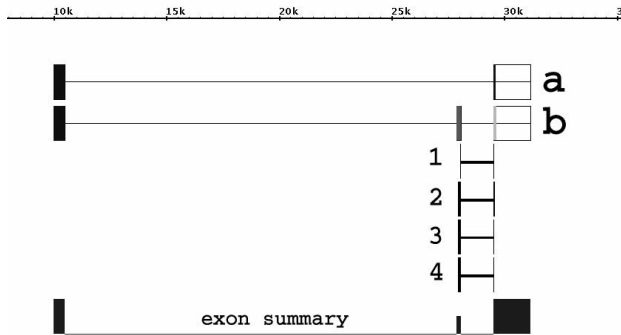


Figure 1. MEOX1 alternative transcripts aligned to genomic sequence. Two alternative transcripts derived from alternative splicing at the MEOX1 locus are shown. One of the transcripts (**b**) contains a 3' homeobox-like motif due to inclusion of an additional exon not present in the other. The first (5'-most) exon of each transcript is translated in the same frame, while the 3' exons are translated in different frames. Thus the two MEOX1 variants share a common amino terminus but differ in the C-terminal regions. Transcripts shown include: (**a**) NM_013999.1 and (**b**) NM_004527.1. InterPro homologies include: (**1**) IPR001356, Homeobox domain (ScanRegExp, PS00027) (**2**) IPR001356, Homeobox domain (HMMSmart, SM00389: HOX) (**3**) IPR001356, Homeobox, (HMMPFam, PF00046: homeobox), and (**4**) IPR001356, Homeobox (ProfileScan, PS0071: HOMEBOX_2).

Key: Transcripts aligned to genomic sequence are shown as rectangles (exons) drawn on top of lines (introns.) The coding regions are filled in and shaded according to translation frame. Exonic regions are visually summarized across all variants in the bottom row as blocks of varying height indicating the number of transcripts that have overlapping sequence at each position.

2.4.2. CD84. CD84 is member of the CD2 family of cell surface molecules and is expressed in numerous blood cell types, including including B-cells, T-cells, monocytes, and platelets [15]. Recently it was shown to enhance cytokine secretion by T-cells by virtue of homophilic binding involving an extra-cellular Ig-like domain that likely permits head-to-head binding between molecules expressed on different cells [15]. This gene is reported to encode several variants, including one, CD84e/s

(accession AF054818.1), which possesses a shortened C-terminal region relative to the other forms [16]. This form (GenBank accession AF054818.1) is shown at the top of Figure 2.

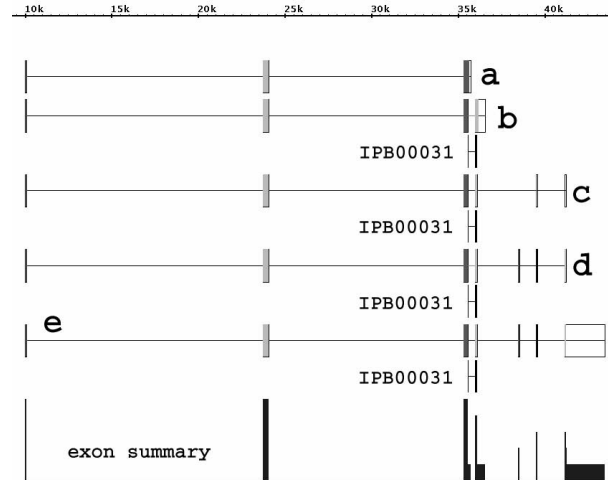


Figure 2. CD84 alternative transcripts aligned to genomic sequence. Alignments between five different CD84 transcript variants and genomic sequence are shown. Proteins encoded by four of the variants (**b-e**) are recognized by Block IPB00031 (Transmembrane 4 family). The top-most transcript (**a**) lacks this domain because of its longer exon 3, which is also the terminal exon in this case. All 5 transcripts contain an immunoglobulin motif aligning to exons 2 and 3. This motif is not shown here since it has the same structure in all variants, which include GenBank accessions (**a**) AF054818.1 (**b**) U96627.1 (**c**) AF054817.1 (**d**) U82988.1 (**e**) NM_003874.1.

The topmost form in Figure 2 lacks a putative transmembrane region that is present in the others, which are recognized by a Blocks "transmembrane 4 family" motif (IPB000301) mapping to exons 3 and 4. Analysis of these sequences using TMHMM, a hidden Markov model-based method useful for recognizing membrane-spanning sequences [17], was done to further refine the location of this putative transmembrane motif (data not shown). According to the TMM-HMM results, the membrane spanning region is located entirely in exon 4 of the CD84 gene within the same region identified by the Blocks IPB000301 profile.

CD84e/s was originally reported as containing the same putative transmembrane motif as the other variants. However, our alignment to genomic sequence, together with protein annotations, show that the truncated protein encoded by isoform AF054818.1 lacks this region and therefore is likely to be secreted. At least one other CD2-like molecule, CD150, is known to have a secreted form [18], and so our interpretation is not unreasonable,

provided the CD84e/s variant reported in Genbank represents a legitimate product of the CD84 locus.

3. Discussion

It is attractive to speculate that alternative transcript structure affecting coding regions generates biologically meaningful functional diversity in the encoded protein products. However, it is possible that in many cases, alternative splicing and other forms of alternative transcript production that affect the coding region actually exert their effects at the level of RNA, such as by differential inclusion of RNA sequences involved in message localization and stability. Furthermore, the diversity of alternative forms could have no biological relevance at all in most cases and may simply be a side effect of RNA processing and whatever as-yet unknown, evolutionary pressures have generated the highly fragmented nature of human genes. Although our results do not resolve these issues, the high incidence of differential motif structure coinciding with alternative transcript structure strongly support the former possibility, since such motifs represent highly conserved regions and therefore are likely to impact function.

We have shown that analyzing protein isoforms with respect to protein motifs is feasible and can be informative whenever the functional or structural significance of the motifs is known. Furthermore, detailed examination of genes producing proteins with distinct motif profiles can lead to intriguing hypotheses as to the functional significance of alternative transcript structure, as described for CD84 and MEOX1.

However, when the functional or structural significance of the affected protein sequence motif is not known, then biological interpretation is difficult or impossible. In light of this, our future efforts will focus on learning how alternative transcript structure affects structural motifs in the encoded proteins; we plan to extend our analysis to include structure-based annotations of protein sequence as described previously [19].

4. Acknowledgements

Ray Wheeler, David Kulp, and Alan Williams wrote genome analysis software that produced pslayout alignments and conceptual translations of genomic sequence. Gang-wu Mei provided InterPro analysis results. We are especially grateful to the BioPerl and Postgresql open source communities for programming libraries and software used in this study.

5. References

- [1] E. S. Lander, *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860-921., 2001.
- [2] J. Xie and D. L. Black, "A CaMK IV responsive RNA element mediates depolarization-induced alternative splicing of ion channels," *Nature*, vol. 410, pp. 936-9., 2001.
- [3] E. Stickeler, F. Kittrell, D. Medina, and S. M. Berget, "Stage-specific changes in SR splicing factors and alternative splicing in mammary tumorigenesis," *Oncogene*, vol. 18, pp. 3574-82., 1999.
- [4] B. Modrek, A. Resch, C. Grasso, and C. Lee, "Genome-wide detection of alternative splicing in expressed sequences of human genes," *Nucleic Acids Res*, vol. 29, pp. 2850-9., 2001.
- [5] J. K. Taylor, Q. Q. Zhang, J. R. Wyatt, and N. M. Dean, "Induction of endogenous Bcl-xS through the control of Bcl-x pre-mRNA splicing by antisense oligonucleotides," *Nat Biotechnol*, vol. 17, pp. 1097-100., 1999.
- [6] D. Schmucker, J. C. Clemens, H. Shu, C. A. Worby, J. Xiao, M. Muda, J. E. Dixon, and S. L. Zipursky, "Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity," *Cell*, vol. 101, pp. 671-84., 2000.
- [7] J. G. Henikoff, S. Pietrokovski, C. M. McCallum, and S. Henikoff, "Blocks-based methods for detecting protein homology," *Electrophoresis*, vol. 21, pp. 1700-6, 2000.
- [8] R. Apweiler, *et al.*, "The InterPro database, an integrated documentation resource for protein families, domains and functional sites," *Nucleic Acids Res*, vol. 29, pp. 37-40., 2001.
- [9] M. Ashburner, *et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, pp. 25-9, 2000.
- [10] K. D. Pruitt and D. R. Maglott, "RefSeq and LocusLink: NCBI gene-centered resources," *Nucleic Acids Res*, vol. 29, pp. 137-40, 2001.
- [11] W. J. Kent and D. Haussler, "Assembly of the working draft of the human genome with gigassembler," *Genome Res*, vol. 11, pp. 1541-8., 2001.
- [12] P. A. Futreal, *et al.*, "Isolation of a diverged homeobox gene, MOX1, from the BRCA1 region on 17q21 by solution hybrid capture," *Hum Mol Genet*, vol. 3, pp. 1359-64., 1994.
- [13] D. Stamatakis, M. Kastrinaki, B. S. Mankoo, V. Pachnis, and D. Karagogeos, "Homeodomain proteins Mox1 and Mox2 associate with Pax1 and Pax3 transcription factors," *FEBS Lett*, vol. 499, pp. 274-8., 2001.
- [14] S. Khorasanizadeh and F. Rastinejad, "Transcription factors: the right combination for the DNA lock," *Curr Biol*, vol. 9, pp. R456-8., 1999.
- [15] M. Martin, X. Romero, M. A. de la Fuente, V. Tovar, N. Zapater, E. Esplugues, P. Pizcueta, J. Bosch, and P. Engel, "CD84 functions as a homophilic adhesion molecule and enhances IFN-gamma secretion: adhesion is mediated by Ig-like domain 1," *J Immunol*, vol. 167, pp. 3668-76., 2001.
- [16] E. Palou, F. Piroto, J. Sole, J. H. Freed, B. Peral, C. Vilardell, R. Vilella, J. Vives, and A. Gaya, "Genomic characterization of CD84 reveals the existence of five isoforms differing in their cytoplasmic domains," *Tissue Antigens*, vol. 55, pp. 118-27., 2000.
- [17] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer, "Predicting transmembrane protein topology with

a hidden Markov model: application to complete genomes," *J Mol Biol*, vol. 305, pp. 567-80., 2001.

[18] J. Punnonen, B. G. Cocks, J. M. Carballido, B. Bennett, D. Peterson, G. Aversa, and J. E. de Vries, "Soluble and membrane-bound forms of signaling lymphocytic activation molecule (SLAM) induce proliferation and Ig synthesis by activated human B lymphocytes," *J Exp Med*, vol. 185, pp. 993-1004., 1997.

[19] M. Cline, G. Liu, A. Loraine, R. Shigeta, J. F. Cheng, G. Mei, D. Kulp, and M. A. Siani-Rose, "Structure-based comparison of four eukaryotic genomes," presented at Pacific Symposium on Biocomputing, Kauai, Hawaii, 2002.