

Reannotation of Human Liver Transcripts From Genome Tiling Microarrays Reveals the Location of Functionally Significant Regions in the Human Genome

Poh Yang Ming, Ali Reza Zamli, Zayed Albertyn and Robert G. Hercus

Application of a unique structured network pattern database platform, SynaBASE™, is presented in the context of novel transcript mapping and rapid genome annotation. 339 novel transcripts from human liver were mapped to the NCBI 35 release of the human genome and subsequently characterized using SwissPROT and simplified protein alphabets in SynaBASE. SynaMap™, a high-throughput mapping application, was used to compute spliced alignments at an average of 146 milliseconds per exon, accuracy of top scoring results were further validated by comparison with BLAT exon positions. Protein function characterization based on conserved family domains and gene ontologies was deduced for 92% of the uncharacterized set using a combination of SynaBASE protein "SIGNIFICANCE" and simplified alphabets of protein motifs. With all analyses being conducted in milliseconds per query on a single CPU server, the results indicate the presence of functionally significant regions in the human genome, which were previously characterized as non-functional.

Introduction

The completion of mammalian genome sequencing projects such as those of human and mouse have significantly enabled a combination of computational and experimental approaches for extensive genome annotation. Creating full catalogues of validated mammalian genes is an essential first step in understanding their biology and is critical in novel drug development. However, the process of comprehensive identification, validation and characterisation of the human genome by conventional means is exhaustive and time consuming. The fact that human genes consist of small coding sequences interspersed with large intronic regions makes the identification and annotation of all transcribed sequences within the entire genome a challenging and as yet incomplete endeavour.

A recent discovery using tiling arrays (a tool for measuring gene expression using sequential overlapping probes), has revealed 339 novel transcripts within the human genome [1]. The availability of an efficient database structure for genomics applications can be useful in enabling rapid

storage and analysis of this novel data. SynaBASE™, a structured network database, is able to store all existing genomics data into a single integrated platform for novel and seminal ultra-high-throughput analysis.

In this paper we have demonstrated the effectiveness of utilising a unique structured network database architecture (SynaBASE) in validating exon positions and confirming the aforementioned novel transcripts. The results are based on the applications of an integrated set of analysis tools that query SynaBASE by mapping transcript data to the genome and functionally characterise raw sequence information using SIGNIFICANCE. SIGNIFICANCE, a unique feature of SynaBASE, is defined as the probability of predicting predecessor and successor characters based on shared sequence motifs in SynaBASE. Processing speeds at each application level is performed in a matter of milliseconds per query [2]. Researchers are therefore able to fully utilise the integrated nature of this system for potentially exhaustive genome annotation and high-throughput genomics.

Materials and Methods

The 339 novel transcripts from the human liver identified by Bertone et. al. (2004) consisted of single exons that were mapped to the Human Genome release 35 (Human genome sequencing consortium) [3-4] using SynaMap™. All SynaMAP alignments were computed at the UCSC Genome Browser database [5].

All novel transcripts were re-annotated with SynaBASE using a combination of SynaMine™ and SynaSearch™ analysis tools. SynaMine was used to mine for pattern SIGNIFICANCE against a SynaBASE version of SwissPROT 43.0 [6] as well as a specialised simple alphabet database compiled from SwissPROT and Smotif™, previously described in [7]. Amino acids are grouped in the simplified alphabet according to their occurrence in columns observed in protein family multiple sequence alignments.

SynaSearch was used for aligning nucleotide transcripts against the SynaBASE version of the SwissPROT database. All applications were run using a single 1.3GHz HP® Itanium™ CPU running Linux.

Results and Discussions

The 339 novel transcripts with a mean exon length of 280 basepairs took 49.5 seconds to search with an average search time of 146 milliseconds per exon. SynaMap was used to report the positions of only the top scoring exons. SynaMap exon positions were further compared to public annotations at the UCSC Human Genome browser database (www.genome.ucsc.edu). SynaSearch nucleotide-protein searches against SynaBASE's Smotif database was also used for find significant protein matches to these exons (Figure 1 below).

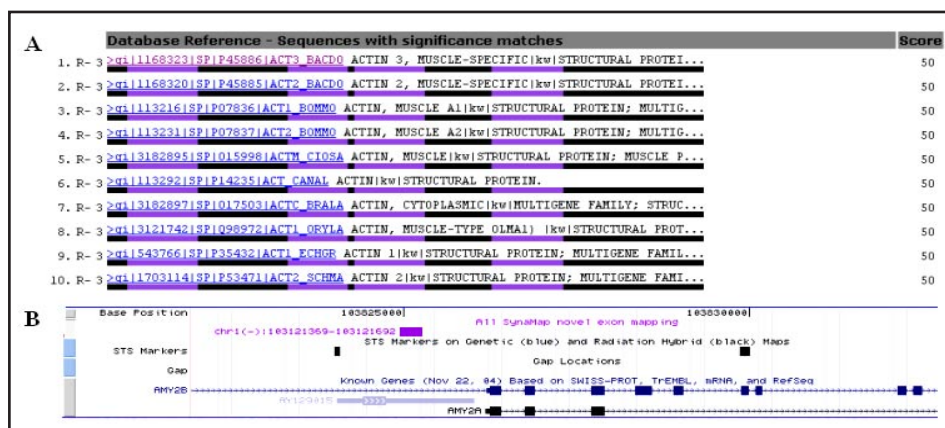


Figure 1: Analysis of SynaMap exon positions for a transcriptional active region on chr1: 103121369-103121692 detected by the tiling microarrays [2]. (A) Selected query showing top ten SynaSearch matching act in based on Smotif database; (B) Display of a SynaSearch match on the UCSC human genome browser database showing alignment with mRNA AY129015 encoding the Q6YL44 protein (hypothetical protein)

Simplified alphabets based on conserved domain motifs are extremely useful in identifying family members in the presence of low sequence similarity. The principle is analogous to scoring matrices such as PAM and BLOSUM [8-9]. SynaBASE Smotif took advantage of this fact to find more significant matches to previously uncharacterized transcripts suspected of belonging to novel genes.

For a number of cases, the findings confirmed that there were few significant matches to known genes when using the SwissPROT database. Most matches found were listed as hypothetical proteins. SynaMine analysis of chr1:103121369-103121692 using Smotif confirmed the match to an actin structural protein (UNIPROT: Q6YL44 annotated as a hypothetical protein - see Figure 2 below) that was earlier found by nucleotide protein searches of the Smotif SynaBASE. This indicates that simplified alphabets, SIGNIFICANCE and SwissPROT keywords can therefore be used for functional annotation in SynaBASE of suspected novel genes.

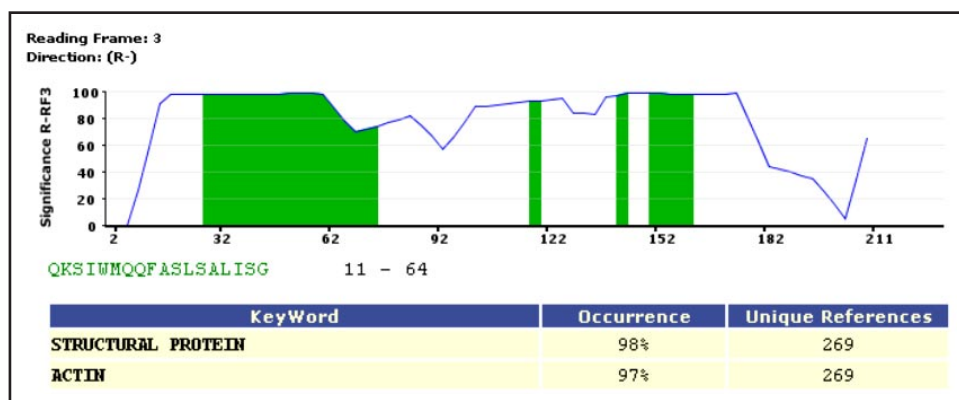


Figure 2: SynaMine analysis of a selected novel transcript region (chr1:103121369-103121692) against SynaBASE's Smotif simplified alphabet database. Pattern SIGNIFICANCE peaks and SwissPROT keywords are shown for the translated 3rd (F3) reverse reading frame.

150 out of the 339 (44%) novel exons found significant matches to SwissPROT proteins in SynaBASE. A number of these could be mapped to gene ontology terms and Interpro protein families describing processes such as apoptosis and protein translation (see Table 1).

However, the Smotif database yielded significantly more results and found 315 out of 339 (93%) matches, some examples of which are listed in Table 2. Therefore Smotif improved search results by almost 50%. In addition SIGNIFICANCE and keywords also correlated with the nucleotide-protein search results (data not shown).

Novel Exon (chromosome: position)	SynaBASE Match (Uniprot Accession)	Interpro ID	Interpro Description	Gene Ontology ID[10]	Gene Ontology Description
chr1: 154156523 - 154156742	O14681	IPR009890	Etoposide-induced 2.4	GO:0012501	programmed cell death
	O14681	IPR009890	Etoposide-induced 2.4	GO:0012502	induction of programmed cell death
	O14681	IPR009890	Etoposide-induced 2.4	GO:0016265	death
chr8: 80531395 - 80532105	Q92901	IPR009000	Translation factor	GO:0044260	cellular macromolecule metabolism
	Q92901	IPR009000	Translation factor	GO:0044267	protein amino acid phosphorylation
	Q92901	IPR009000	Translation factor	GO:0050875	cellular physiological process
chr4: 125663955 - 125664266	Q9NZ01	IPR001104	3-oxo-5-alpha-steroid 4-dehydrogenase, C-terminal	GO:0005623	cell
	Q9NZ01	IPR001104	3-oxo-5-alpha-steroid 4-dehydrogenase, C-terminal	GO:0016020	membrane
	Q9NZ01	IPR001104	3-oxo-5-alpha-steroid 4-dehydrogenase, C-terminal	GO:0016021	integral to membrane

Table 1: SynaSearch annotation results for selected novel transcribed regions.

Novel Exon (chromosome:position)	SynaBASE Match (UniProt Accession)	InterPro Description
chr2 (-):176341969-176342190	O04928	Phosphatidate cytidyltransferase
chr15 (-):36328158-36328377	O08912	Ricin B lectin
chr15 (-):36328158-36328377	O08912	Glycosyl transferase, family 2
chr15 (-):36328158-36328377	O08912	Ricin B-related lectin
chr6 (-):141090672-141090891	O08999	Aspartic acid and asparagine hydroxylation site
chr6 (-):141090672-141090891	O08999	EGF-like, subtype 2
chr6 (-):141090672-141090891	O08999	EGF-like calcium-binding
chr6 (-):141090672-141090891	O08999	Aldehyde dehydrogenase
chr6 (-):141090672-141090891	O08999	Matrix fibril-associated
chr12 (+):80786733-80787007	O09159	Glycoside hydrolase, family 38
chr12 (+):80786733-80787007	O09159	Galactose mutarotase-like

Table 2: Selected novel exons that were missed by SwissPROT but gave matches using Smotif (filtered with a major alignment cut-off score of 15).

Conclusion

Transcribed regions lying outside of previously annotated genes are expected to correspond primarily to unannotated exons alternative splicing events; underrepresented 3' and 5' untranslated regions; non protein coding RNA transcripts, and novel transcripts coding for functional proteins [2]. The results show a combination of standard and novel sequence analysis applications using the SynaBASE architecture to functionally annotate these novel transcript data. Accurate transcript-to-genome could be accomplished in a fraction of the time taken by similar methods on more powerful architectures. Furthermore the application of specialised Smotif database for remote homolog detection showed improvements in finding results by up to 50% when compared to a normal protein database search. In addition to qualitative characterization, all applications built around SynaBASE are exceptionally efficient in terms of speed and accuracy.

The data presented here were generated on modest computing resources in a fraction of the time taken by large CPU clusters. SynaBASE could be used to perform similar analysis resulting from multiple microarray or other large-scale expression studies in a high-throughput environment for functional annotation of newly identified transcriptionally active genomic DNA.

REFERENCES

1. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M. Global identification of human transcribed sequences with genome tiling arrays. *Science* (2004) 306(5705):2242-6.
2. Albertyn, Z.I., Wong CS, Tay, LC, Ramachandran M, Hercus RG. A new Approach to Genome-wide Annotation Based upon Calculation of Significance from a Structured Pattern Database. *Synamatix* June 2004 MTN. Available from www.synamatix.com.
3. International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409:860-921.
4. Istrail S, Sutton GG, Florea L, Halpern AL, Mobarry CM, Lippert R, Walenz B, Shatkay H, Dew I, Miller JR, Flanigan MJ, Edwards NJ, Bolanos R, Fasulo D, Halldorsson BV, Hannenhalli S, Turner R, Yooseph S, Lu F, Nusskern DR, Shue BC, Zheng XH, Zhong F, Delcher AL, Huson DH, Kravitz SA, Mouchard L, Reinert K, Remington KA, Clark AG, Waterman MS, Eichler EE, Adams MD, Hunkapiller MW, Myers EW, Venter JC. Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc Natl Acad Sci U S A.* (2004) 101(7):1916-21
5. Kent, W.J., Sugnet, C. W., Furey, T. S., Roskin, K.M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002).The Genome Browser database at UCSC. *Genome Res.* 12:996-1006
6. Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S. and Schneider M. The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* (2003) 31:365-370.
7. Albertyn, Z.I., Tay, L.C., Ramachandran M, Hercus R.G., Application of Novel Structured Pattern Database to Identifying Sequence Motif Conservation in Protein Families. *Synamatix* October 2004 MTN Available from www.synamatix.com.
8. Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. (1978) A model of evolutionary change in proteins. matrices for detecting distant relationships In M. O. Dayhoff, (ed.), *Atlas of protein sequence and structure*, volume 5, pp. 345-358 National biomedical research foundation Washington DC
9. Henikoff, S., and Henikoff, J.G. (1992) Amino acid substitution matrices from protein. *Proc. Nat. Acad. Sci. USA* 89(22):10915-10919
10. The Gene Ontology Consortium (2000). Gene Ontology: tool for the unification of biology. *Nature Genet.* 25: 25-29

This Application Note is for distribution to Synamatix members, associate members and mailing list subscribers only. The contents are provided for personal, non-commercial purposes only and are protected by various national and international intellectual property laws, conventions and treaties. All title and intellectual property rights in and to Synamatix, SynaBASE, SynaMine, and SynaSuite and the accompanying printed materials are owned by Synamatix sdn bhd. Other trademarks or names are used only in an editorial fashion and to the benefit of the respective trademark owner with no intention of the infringement of the trademark. All trademarks or service marks are the property of their respective owners.