

# Splicing Profile Based Protein Categorization between Human and Mouse Genomes by Use of the DDBJ Web Services

Åke Västermark<sup>1,2</sup>      Yasumasa Shigemoto<sup>3</sup>  
vasterma@stats.ox.ac.uk      yshigemo@genes.nig.ac.jp  
Takashi Abe<sup>1</sup>      Hideaki Sugawara<sup>1</sup>  
takaabe@genes.nig.ac.jp      hsugawar@genes.nig.ac.jp

- <sup>1</sup> Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics and the Graduate University for Advanced Studies, Mishima, Shizuoka 411-8540, Japan  
<sup>2</sup> Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK  
<sup>3</sup> Life Science Systems Division, Fujitsu Limited, Shinkamata, Tokyo-to 144-8588, Japan

## Abstract

In one scenario of gene evolution, exon shuffling plays a fundamental role in increasing gene diversity. This paper is an appraisal of the biological relevance of categorising proteins by their splicing profiles (exon-intron structures). The central question is whether protein function is more correlated with splicing profiles than sequence similarity, or not. To approach this question, a splicing profile similarity (SPS) index, which measures relative exon length discrepancy, was devised. Arbitrary human proteins were compared, in terms of SPS and amino acid sequence similarity, to their 1) mouse orthologues and 2) human paralogues, which epitomise functional equivalence and non-equivalence, respectively, to methodically elucidate the global relationship between a) biological function, b) splicing profile similarity, and c) sequence similarity. Protein function is more correlated with splicing profile similarity than sequence similarity as demonstrated by the fact that human-mouse orthologues (HMOs) display significantly higher splicing profile similarity than do human-human paralogues (HHPs), despite the mutual sequence similarity between these two categories. This finding indicates that splicing profile-based protein categorisation is biologically meaningful.

**Keywords:** splicing profile similarity, exon-intron structures, human-mouse orthologues, human-human paralogues, web services

## 1 Introduction

Proteins can be categorised by various criteria. Normally, it is desirable to categorise proteins by their biological functions. Because it is a difficult undertaking to assign a given protein sequence to a known function, however, proteins are often categorised by sequence similarity instead [6, 10, 11]. Nevertheless, there are many alignment-independent, alternative ways of classifying proteins. For example, it is feasible to classify G-protein coupled receptors (GPCRs) through different statistical analyses of their primary amino acid sequences, including: 1) n-tuplet composition studies [2], and 2) principal physicochemical characteristics studies [12]. More advanced, alignment-independent efforts include: 3) examination of the distances between conserved, terminal and transmembrane, key residues [14], and 4) examination of the lengths of extracellular and intracellular loops and terminals [15]. The above methods (1.4) all give rise to GPCR classifications largely coincidental with the conventional sequence-

and pharmacology-based classifications, confirming that these alignment-independent strategies are biologically germane.

A related theme, there are many examples of successful efforts to include structural information in homology searches, such as: 1) protein fold recognition methods that take secondary structure predictions into account [5, 7, 17, 18], and 2) membrane protein detection optimised Smith-Waterman and profile algorithms that incorporate topology forecasts [8]. In these cases, the inclusion of structural information increases both the sensitivity and the specificity of homology searches, again confirming the usefulness of alignment-independent information. This article is an assessment of the value of categorising proteins by their exon-intron structures (alias splicing profiles), another alignment-independent feature, which is, to my knowledge, yet unexplored in the protein classification context.

There are two divisive theories of the origin and evolution of exon-intron structures: 1) the exon theory of genes, also known as the introns early theory, and 2) the insertional theory of introns, also known as the introns late theory. First, the exon theory of genes suggests that the primordial genes contained introns, and evolved through exon shuffling, the process by which discrete sequences encoding stably folding building blocks and functional domains are rearranged to create new proteins. Second, the insertional theory of introns advocates that the archaic genome was virtually intron-free, and that the exon-intron structures of modern genes are merely random manifestations of undirected transposition. Proteome-wide, statistical analyses of intron phases, i.e. the location of splice sites relative to the pertinent domain boundaries and reading frames, indicate that both sides of the argument are partially correct; while some introns are, apparently, randomly placed, there is also a statistically significant proportion of splice sites that coincide with protein domain boundaries and reading frames [3, 4, 9, 13, 16]. Furthermore, the conservation of exon-intron structures, within clusters of functionally interrelated domains, is irregular; certain extracellular-signalling and nuclear domains consistently display conservation of their splicing profiles greater than the conservation of their sequence homologies, whereas many subgroups of intracellular-signalling domains always exhibit the reverse pattern [1].

Of course, this preamble raises the greater question: what is the global relationship between 1) protein function, 2) sequence similarity, and 3) splicing profile similarity? This paper approaches this question by introducing the splicing profile similarity (SPS) index, a new way of comparing the resemblance of the exon-intron structures of a protein pair. Here, the central assumption is that human-mouse orthologues (HMOs) and humanhuman paralogues (HHPs) epitomise functional equivalence and non-equivalence, correspondingly. Inasmuch as this conjecture is true, it is possible to evaluate the degree of correlation between function and SPS/homology. The results of this experiment suggest that protein function is, generally, more correlated with splicing profiles than sequence similarity.

This project relies on a DNA databank of Japan (DDBJ) workflow, which takes advantage of the DDBJ web services [19]. Because this project was enabled by the web services of the DNA databank of Japan, this paper will also provide an introduction to the DDBJ web services.

## 2 Methods

### 2.1 The Splicing Profile Similarity (SPS) Index

The splicing profile similarity (SPS) index, which is only applicable to same exon count protein pairs, is a measure of exon-intron structure resemblance:

$$SPS = 1 / \left( \sum \frac{|e1 - e2|}{e1 + e2} + 1 \right),$$

where  $e1$  and  $e2$  are the lengths (in nucleotides) of parallel exons. This index is relative, because the absolute exon length difference ( $|e1 - e2|$ ) is divided by the total exon length sum ( $e1 + e2$ ); therefore this index is more sensitive to short exon divergence, and vice versa, less sensitive to long

exon divergence. The number 1 is divided by (the sum of the relative exon length differences + 1 (to prevent division by zero)). This means that SPS scores range from infinitesimal, for infinitely dissimilar protein pairs, to 1, for identical pairs. This index offers an alternative measure of protein similitude, which can be contrasted to measures of sequence similarity, such as BLAST E-value.

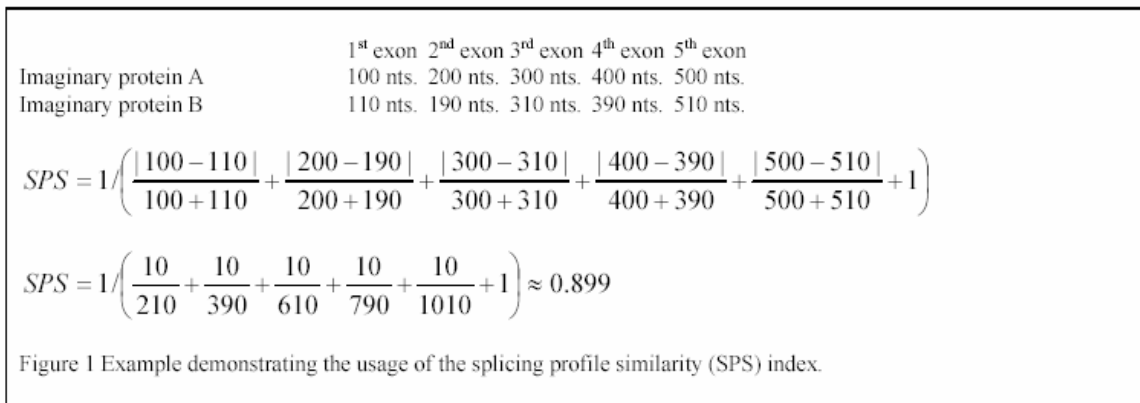


Figure 1: Example demonstrating the usage of the splicing profile similarity (SPS) index.

## 2.2 The Splicing Profile Similarity (SPS) Work Flow

This workflow, which is outlined in Figure 2, was enabled by, and implemented through, the DNA data-bank of Japan (DDBJ) web services (<http://www.xml.nig.ac.jp/>) [19]. The web services provide programmatic interfaces to databases and data analytical tools and make efficient system development possible. To systematically elucidate the relationship between function, sequence similarity and splicing profile similarity, arbitrary human ENSEMBL proteins are compared, in terms of SPS and sequence similarity (BLAST E-value), to their A) mouse orthologues, and B) closest human paralogues, respectively.

In this study, human-mouse orthologues (HMOs) are defined as pairs of human and mouse proteins that have synonymous ENSEMBL identifiers [20]. According to this nomenclature-based definition of orthology, there are 8161 orthologous gene pairs in the 21.34d and 21.32d releases of the human and mouse ENSEMBL databases. Because some genes give rise to multiple transcripts, there is a total of 15266 orthologous relationships on the transcript level, between the human and mouse ENSEMBL transcript databases. From the set of 15266 orthologous transcript relationships, only a subset of 8270 transcripts display the same number of exons. Furthermore, in the set of 15266 orthologous pairs, only 4646 transcripts are not the result of alternative splicing. The intersection between the same-exon-count and non-alternative-splicing subsets is sized 3438, and was used as the final HMO set. The sequence identity between the HMO pairs was determined using BLAST.

Closest human-human paralogues (HHPs) are defined as pairs of human proteins that have reciprocal sequence similarity greater than the sequence similarity between either of the proteins and any other human ENSEMBL protein, and greater sequence similarity than a certain threshold. According to this definition, there are 507,354 paralogous protein pairs and 137,644 paralogous gene pairs, and subsets of 131,630 and 74,847 paralogous protein and gene pairs display the same number of exons. A subset of 57,604 paralogous protein (gene) pairs do not derive from alternative splicing. The intersection between the same-exon-count subset and the no-alternative-splicing subset contains 37,105 paralogous genes. Because the HHP data set is much larger than the HMO data set, a randomly selected, HMO-data-set-sized subset of the HHP data set is used instead of the full HHP data set.

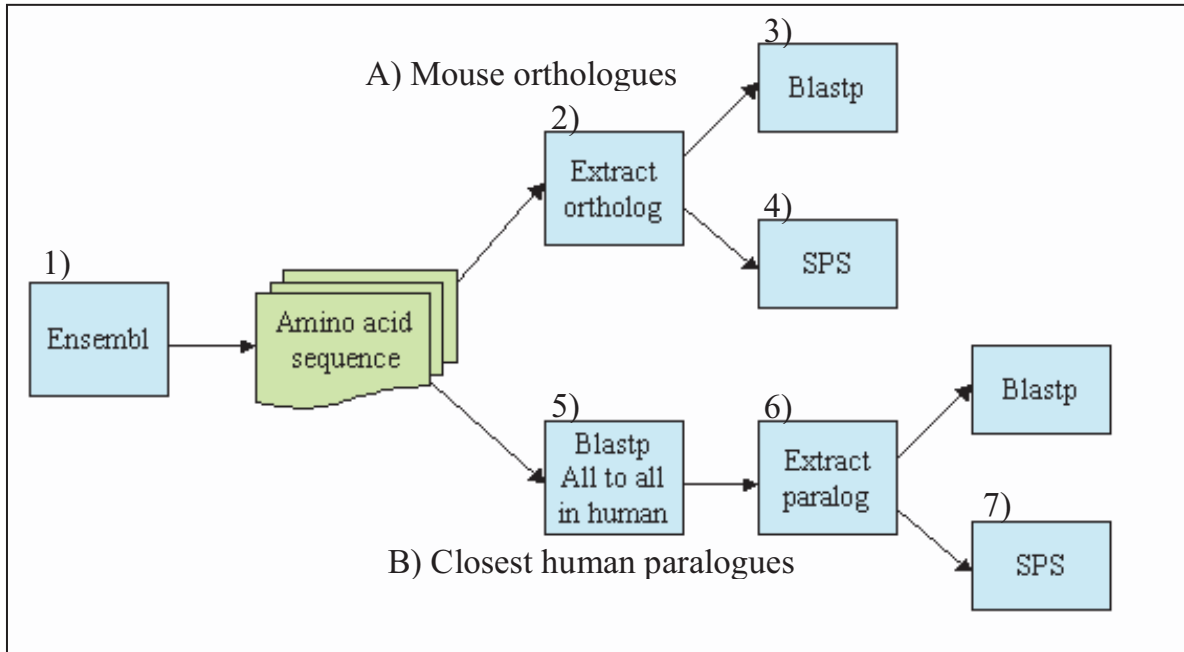


Figure 2: Flowchart of the splicing profile similarity (SPS) workflow. SPS workflow details: 1) retrieve protein information, including human and mouse amino acid sequences, genomic locations and gene symbols, from the ENSEMBL database, 2) extract orthologous genes from the data set, 3) determine the sequence similarity between the orthologous genes using BLASTP, 4) calculate splicing profile scores for all orthologous proteins, 5) perform all-to-all BLASTP homology search of all human proteins, 6) extract paralogous protein pairs, and 7) calculate the splicing profile scores for paralogous protein pairs.

### 3 Results

The human-mouse orthologue and human-human paralogue plots of splicing profile similarity (SPS) against sequence similarity are summarized in Figure 3(a) and Figure 3(b) respectively. The difference of the distribution of the dots in the figures demonstrate that protein function is more correlated with splicing profile similarity than sequence similarity.

### 4 Discussion

Protein function is more correlated with splicing profile similarity than sequence similarity as demonstrated by the fact that human-mouse orthologues (HMOs) display significantly higher splicing profile similarity than do human-human paralogues (HHPs), despite the mutual sequence similarity between these two categories.

This finding depends on multiple suppositions, which are open to discussion:

- The definition of HMOs as pairs of human and mouse proteins that have cognate ENSEMBL identifiers, and the definition of HHPs as pairs of human proteins that have reciprocal sequence similarity greater than the sequence similarity between either of the proteins and any other human ENSEMBL protein

HMOs are most reliably defined as pairs of human and mouse proteins that have reciprocal sequence similarity greater than ((the sequence similarity between the human protein and any other mouse protein) and (the sequence similarity between the mouse protein and any other human protein)), because there are isolated examples of ENSEMBL synonyms that are not orthologues and, vice versa, orthologues that have dissimilar ENSEMBL names. Nevertheless, human-mouse ENSEMBL synonyms

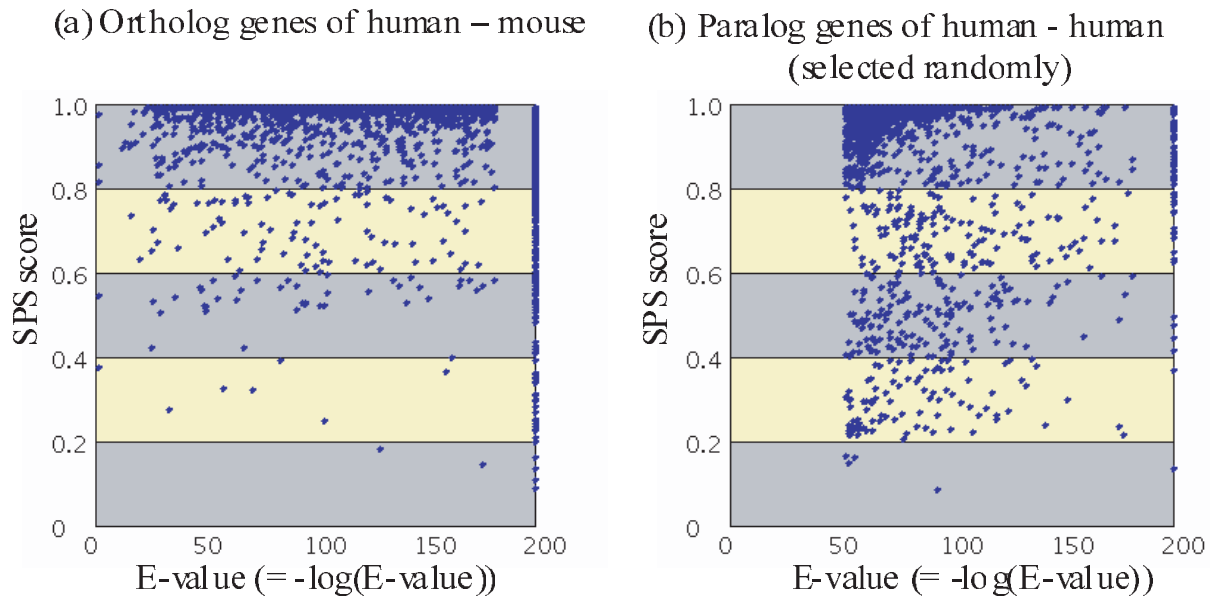


Figure 3: The human-mouse orthologue (a) and human-human paralogue (b) plots of splicing profile similarity (SPS) scores against amino acid sequence similarity. Because the HHP data set is much larger than the HMO data set, a randomly selected, HMO-data-set-sized subset of the HHP data set is used instead of the full HHP data set. Raw E-values are plotted as negative logarithms of the raw E-values, and zero raw E-values are plotted as 200. The gap between 175 and 200 on in the E-value dimension is due to an exponent rounding property of BLAST that rounds exponents greater than 175 to zero. The significance of the plots is explained through a numerical summary in Table 1.

Table 1:

Table 1a)

	SPS<0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1	sum	ratio
BLAST e-175 - 0	1	4	11	13	10	46	39	77	184	1635	2020	0.81
e-150 - e-175	0	0	0	1	1	5	6	8	16	223	260	0.86
e-125 - e-150	0	1	0	0	0	5	6	4	20	236	272	0.87
e-100 - e-125	0	0	1	0	0	4	8	9	23	241	286	0.84
e-75 - e-100	0	0	0	1	0	7	9	8	16	193	234	0.82
e-50 - e-75	0	0	0	2	1	6	4	13	8	166	200	0.83
e-25 - e-50	0	0	1	0	0	8	2	9	22	100	142	0.7
0 - e-25	0	0	0	1	1	1	2	1	2	16	24	0.67

Table 1b)

	SPS<0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1
BLAST e-175 - 0	45.0	34.3	15.7	13.4	18.0	14.3	14.6	16.3	14.9	10.1
e-150 - e-175	0	0	0	9.0	8.0	9.2	7.8	7.0	8.8	5.3
e-125 - e-150	0	11.0	0	0	0	5.6	7.7	6.8	6.2	5.0
e-100 - e-125	0	0	7.0	0	0	6.5	5.5	6.9	5.2	4.5
e-75 - e-100	0	0	0	6.0	0	4.9	5.0	5.5	5.3	4.2
e-50 - e-75	0	0	0	5.0	5.0	4.7	3.5	4.2	4.6	3.2
e-25 - e-50	0	0	15.0	0	0	4.8	2.5	2.9	3.3	2.8
0 - e-25	0	0	0	4.0	23.0	2.0	4.5	27.0	3.5	2.4

Table 2:

Table 2a)

	SPS<0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1	sum	ratio
BLAST e-175 - 0	1	7	13	12	17	21	42	81	128	496	818	0.61
e-150 - e-175	1	1	5	13	9	17	40	41	54	207	388	0.53
e-125 - e-150	0	0	14	31	24	67	70	76	94	283	659	0.43
e-100 - e-125	0	3	44	33	85	151	112	100	160	637	1325	0.48
e-75 - e-100	1	4	53	93	153	187	170	189	219	4307	5376	0.8
e-50 - e-75	0	23	167	144	195	172	162	221	831	26624	28539	0.93
e-25 - e-50	0	0	0	0	0	0	0	0	0	0	0	0
0 - e-25	0	0	0	0	0	0	0	0	0	0	0	0

Table 2b)

	SPS<0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1
BLAST e-175 - 0	52.0	25.6	13.8	9.7	21.4	16.9	15.6	12.9	10.8	5.3
e-150 - e-175	40.0	23.0	9.8	9.0	10.7	5.5	7.8	8.5	7.5	3.0
e-125 - e-150	0	0	8.9	9.9	5.3	3.9	6.3	7.0	6.0	2.3
e-100 - e-125	0	18.3	6.2	5.8	5.9	4.8	6.0	5.7	4.0	1.4
e-75 - e-100	39.0	19.8	8.7	5.4	4.2	4.1	4.4	3.2	2.7	1.1
e-50 - e-75	0	20.8	8.9	4.9	3.6	3.9	3.7	2.6	1.4	1.0
e-25 - e-50	0	0	0	0	0	0	0	0	0	0
0 - e-25	0	0	0	0	0	0	0	0	0	0

Tables (1 a,b and 2 a,b) showing (orthologue and paralogue) statistics, comprising (counts, and exon count averages). The tables constitute a tabular summary of the data in the human-mouse orthologue and human-human paralogue plots. The information has been divided between 8 separate BLAST E-value bands and 10 separate SPS score bands, to facilitate data interpretation. In general, both the orthologues and the paralogues, which represent functional equivalence and non-equivalence, respectively, display a wide range of E-values. The result that orthologues display significantly higher splicing profile similarity than paralogues, which are comparable in terms of sequence similarity, do, suggests that protein function is more correlated with splicing profiles than sequence similarity, given the central assumption that HMOs and HHPs embody functional equivalence and non-equivalence, respectively. The accumulation of high SPS and low homology data points in the paralogue data is probably an artifact deriving from the fact that the high SPS and low homology paralogue data points have an unusually low exon count (close to 1).

are virtually always orthologues. This means that the ENSEMBL nomenclature offers a generally accepted, sequence similarity-independent, way of establishing orthology. In fact, the definition of HMOs as pairs of human and mouse proteins that have synonymous ENSEMBL identifiers opens up the prospect of measuring the degree of correlation between protein function and SPS/homology, precisely because it is sequence similarity-independent. That is, a human-mouse orthologue definition founded on sequence similarity would not permit assessment of the function-SPS and function-homology correlations. Accordingly, the nomenclature-based orthologue definition is both reliable and necessary.

With regard to the definition of HHPs, it agrees with the traditional homology-based one. While this definition is not universally applicable, it suffices, because the paralogues merely serve as benchmarks in this study. That is, it is not imperative whether the alleged paralogues are true paralogues or not; it is only important that they are not orthologues.

- The assumption that HMOs and HHPs represent functional equivalence and non-equivalence, respectively

Given that the relevant orthologue and paralogue definitions are practically adequate, it is logical to presume that the data points on the HMO and HHP plots epitomise functional equivalence and non-equivalence, respectively.

- The assumption that the SPS index and amino acid sequence similarity adequately reflect actual splicing profile similarity and sequence similarity, respectively

The SPS index does indeed mirror splicing profile similarity, because it is essentially a measure of exon length discrepancy. Nevertheless, the splicing profile similarity index is, ultimately, arbitrary, as the SPS equation could just as well be formulated in many other ways. Furthermore, because the SPS index is only applicable to same exon count protein pairs, it unfortunately discriminates against different exon count protein pairs with otherwise highly similar exon-intron structures. It should also be noted that since transcriptional start site detection is relatively unreliable, there is a risk that length discrepancy in leading exons may dominate the SPS values.

## 5 Concluding Remarks and Project Website

In conclusion, the finding that protein function is, generally, more correlated with splicing profile similarity (SPS) than sequence similarity implies that exon-intron structures are biologically germane, and buttresses the exon theory of genes. This result indicates that splicing profile-based protein categorisation is biologically meaningful.

It is to be noted that the work flow for the study of SPS was efficiently accomplished thanks to the DDBJ web services. Web services offer programmatic interfaces for the development of flexible and seamless integration of bioinformatics resources that are distributed in the Internet. We are able to expect that rich work flows composed of multiple web services for various research aims in bioinformatics will be also available in the public domain in addition to individual web services.

Information about this project, including a flowchart and all of the results, is available online at the Biportal website of the DNA databank of Japan (<http://bioportal.ddbj.nig.ac.jp/sps/index.html>).

## Acknowledgments

This project was jointly supported by the program of advancement and standardization of bioinformatics database by Institute for Bioinformatics Research and Development of the Japan Science and Technology (JST) agency, and the “Research and Development of Biological Portal Sites of the New Generation” scheme of the “Special Coordination Funds for Promoting Science and Technology” program of the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT). Åke Västermark’s work was also supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT).

## References

- [1] Betts, M.J., Guigo, R., Agarwal, P., and Russell, R.B., Exon structure conservation despite low sequence similarity: A relic of dramatic events in evolution?, *EMBO J.*, 20:5354–5360, 2001.
- [2] Daeyaert, F., Moereels, H., and Lewi, P.J., Classification and identification of proteins by means of common and specific amino acid n-tuples in unaligned sequences, *Comput. Methods Programs Biomed.*, 56:221–233, 1998.
- [3] de Souza, S.J., Long, M., Klein, R.J., Roy, S., Lin, S., and Gilbert, W., Toward a resolution of the introns early/late debate: Only phase zero introns are correlated with the structure of ancient proteins, *Proc. Natl. Acad. Sci. USA*, 95:5094–5099, 1998.
- [4] Fedorov, A., Fedorova, L., Starshenko, V., Filatov, V., and Grigor’ev, E., Influence of exon duplication on intron and exon phase distribution, *J. Mol. Evol.*, 46:263–271, 1998.

- [5] Fischer, D. and Eisenberg, D., Protein fold recognition using sequence-derived predictions, *Prot. Sci.*, 5:947–955, 1996.
- [6] Graul, R.C. and Sadee, W., Evolutionary relationships among G protein-coupled receptors using a clustered database approach, *AAPS PharmSci.*, 3:E12, 2001.
- [7] Hargbo, J. and Elofsson, A., Hidden Markov models that use predicted secondary structure for fold recognition, *Proteins*, 36:68–76, 1999.
- [8] Hedman, M., Deloof, H., von Heijne, G., and Elofsson, A., Improved detection of homologous membrane proteins by inclusion of information from topology predictions, *Protein Sci.*, 11:652–658, 2002.
- [9] Kaessmann, H., Zollner, S., Nekrutenko, A., and Li, W.H., Signatures of domain shuffling in the human genome, *Genome Res.*, 12:1642–1650, 2002.
- [10] Karchin, R., Karplus, K., and Haussler, D., Classifying G-protein coupled receptors with support vector machines, *Bioinfo.*, 18:147–159, 2002.
- [11] Kuipers, W., Oliveira, L., Vriend, G., and Ijzerman, A.P., Identification of class-determining residues in G protein-coupled receptors by sequence analysis, *Receptors Channels*, 5:159–174, 1997.
- [12] Lapinsh, M., Gutcaits, A., Prusis, P., Post, C., Lundstedt, T., and Wikberg, J.E., Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences, *Protein Sci.*, 11:795–805, 2002.
- [13] Long, M., Rosenberg, C., and Gilbert, W., Intron phase correlations and the evolution of the intron/exon structure of genes, *Proc. Natl. Acad. Sci. USA*, 92:12495–12499, 1995.
- [14] Moereels, H., Lewi, P.J., Daeyaert, F., Schenck, E., and Janssen, P.A., The alpha and omega of G-protein coupled receptors: A novel method for classification. Part 2. Bin classification, *Receptors Channels*, 5:139–148, 1997.
- [15] Otaki, J.M. and Firestein, S., Length analyses of mammalian G-protein-coupled receptors, *J. Theor. Biol.*, 211:77–100, 2001.
- [16] Patthy, L., Genome evolution and the evolution of exon-shuffling, *Gene*, 238:103–114, 1999.
- [17] Rice, D. and Eisenberg, D., A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence, *J. Mol. Biol.*, 267:1026–1038, 1997.
- [18] Rost, B., Schneider, R., and Sander, C., Protein fold recognition by prediction-based threading, *J. Mol. Biol.*, 270:471–480, 1997.
- [19] Sugawara, H. and Miyazaki, S., Biological SOAP servers and web services provided by the public sequence data bank, *Nucleic Acids Res.*, 31(2):3836–3839, 2003.
- [20] <http://www.ensembl.org/>