Identification of embryo specific human isoforms using a database of predicted alternative splice forms

Heike Pospisil

Center for Bioinformatics, University of Hamburg, Bundesstrasse 43, D-20146 Hamburg, Germany, pospisil@zbh.uni-hamburg.de, phone: +49-40-42838-7303, fax: +49-40-42838-7312

Abstract

Alternative splicing is one of the most important mechanisms to generate a large number of mRNA and protein isoforms from a small number of genes. Its study became one of the hot topics in computational genome analysis. The repository EASED (Extended Alternatively Spliced EST Database, http://eased.bioinf.mdc-berlin.de/) stores a large collection of splice variants predicted from comparing the human genome against EST databases. It enables finding new unpublished splice forms that could be interesting candidate genes for stage specific, diseases specific or tissue specific splicing. The main idea behind selecting specific splice forms is to compare the number and the origin of ESTs confirming one isoform with the number and the origin of ESTs confirming the opposite isoform. A measure asDcs is introduced to take into account the unequal distribution of ESTs per splice site on one hand, and the possible uncertainties due to the relatively low quality of EST data on the other hand. First, the number of ESTs per splice site is scaled with a modified Hill function. The measure asDcs computes in the second step the distance of each pair of ESTs from equipartition. Equipartition exists if for every number of adult ESTs the same number of embryonic ESTs. The effect of several input parameters for the scaling of true EST values is analysed and can be reproduced on http://cardigan.zbh.uni-hamburg.de/asDcs. Some of the obtained best scoring hits for selected parameters (*transcription factor P65*, *drebrin*, *and fetuin*) have been already described in literature as been involved in embryonic development. This result shows the plausibility of this approach and looks promising for the identification of unplublished embryo specific alternative splice sites in human.

1 Introduction

Due to the enormous increase of available sequence data, the study of alternatively spliced transcripts became one of the hot topics in computational genome analysis. Alternative Splicing can modulate transcript expression levels either by subjecting mRNAs to the nonsense mediate decay (NMD) or by altering the structure of the gene product by inserting or deleting protein parts. Thereby, alternative splicing is one of the most important mechanisms to generate a large number of mRNA and protein isoforms from a small number of genes. The estimation of frequency of alternatively spliced human transcripts ranges from 35% to 79% [Mironov et al. 1999, Brett et al. 2000, Kan et al. 2001, Modrek et al. 2001, Johnson et al. 2003].

Several databases predicting altered isoforms have been developed during the last years: e.g. ASAP [Lee et al. 2003], ASD [Thanaraj et al. 2004], ASDB [Dralyuk et al. 2000], ECgene [Kim et al. 2005], SpliceDB [Burset et al. 2001], and SpliceNest [Krause et al. 2002]. These

databases (amongst others) provide information about splice variants of distinct genes. The majority of these resources predict splice variants on the basis of expressed sequence tags (ESTs). Since EST sequences are often of poor quality, they have to be carefully checked before using them. Nevertheless, by now, ESTs have the largest potential to examine the expression status of a cell and the occurrence of alternative splicing in distinct tissues, developmental stages or diseases. Therefore, the repository EASED (Extended Alternatively Spliced EST Database) [Pospisil et al. 2004] stores EST-based predicted splice variants including all available EST information. EASED holds histological, pathological and developmental information from ESTs both for alternative and for the corresponding so called normal (or constitutive) splice form. In contrast to the above mentioned databases, EASED uses flexible queries to select interesting genes according to, e.g. histological, pathological and developmental information. This enables to find new unpublished splice forms that could be interesting candidate genes for stage specific, diseases specific or tissue specific splicing without knowing any database identifier or gene name.

The main idea behind selecting specific splice forms is to compare the number and the origin of ESTs confirming one isoform with the number and the origin of ESTs confirming the opposite isoform. Besides the problem of varying sequence and annotation quality, a further difficulty is often neglected: different histological, developmental or pathological original sources are often over- or underrepresented in the presently available EST libraries. In particular, there exists a clear bias towards cancer ESTs and adult ESTs. For this reason, a measure is required that takes into account the quality problem as well as the bias problem. Here, we present such a measure and apply it to the selection of human embryo specific isoforms.

2 Methods

2.1 Prediction of alternative splice sites

The algorithm used to identify alternative splice forms takes the currently available genomic sequences (Ensembl [Birney et al. 2004] release 19, with 31,609 human transcripts) and aligns these sequences to all available ESTs (dbEST release from February 2005 with 25,707,314 human ESTs) using WU-Blast [Gish 2004] and the sim4 program [Florea et al. 1998]. The transcripts are beforehand masked by MaskerAid [Bedell et al. 2000]. A matching pair of genomic sequence and EST has to show at least two colinear high scoring pairs. A splice variant may then be identified if we can find aligned ESTs revealing variations (deletion/insertion) that suggest alternative splicing. In the following, the transcript referring to the Ensembl database are indicated as the ‘constitutive form‘ or ‘normal form’. In contrast, the EST based predicted isoform is named ‘alternative form’. ESTs representing the normal form are indicated ‘constitutively spliced ESTs‘ (csESTs) and ESTs representing the opposite, alternative form are named ‘alternatively spliced ESTs‘ (asESTs).

2.1.1 EASED: database and web interface

The *Extended Alternatively Spliced EST Database* (EASED) [Pospisil et al. 2004] project is establishing a comprehensive database of alternatively spliced human mRNAs. Moreover, EASED includes useful biological information and provides the possibility to search for biologically relevant data and for candidate genes for the origin of diseases. All processed and annotated alternative splice forms are stored in a MySQL database. It allows analysing each splice site separately instead of analysing complete transcripts.

An online version of EASED comprises a selection of possibly interesting MySQL queries and is available at http://eased.bioinf.mdc-berlin.de/.

2.2 The new distance measure asDcs

To predict specific splice sites, the corresponding asESTs and csESTs are compared accoring to a predefined classification. Such a classification distinguishes two categories *c1* and *c2*, e.g. (i) ‘tumor‘ and ‘normal‘ or (ii) ‘tissue:liver‘ and ‘tissue:all except for liver‘ or (iii) ‘embryo‘ and ‘adult‘. Each EST thus has a splice type (*as* or *cs*) and is assigned to one of the two classification *c1* and *c2*. Thus, we divide the set of all ESTs into four different sets and define corresponding counts:

$$ec(as,c1) = \left|EST_{as} \cap EST_{c1}\right|$$
$$ec(as,c2) = \left|EST_{as} \cap EST_{c2}\right|$$
$$ec(cs,c1) = \left|EST_{cs} \cap EST_{c1}\right|$$
$$ec(cs,c2) = \left|EST_{cs} \cap EST_{c2}\right|$$

where EST_{c1} is the set of all ESTs falling into category *c1*, and EST_{c2} is the set of all ESTs falling into category *c2* (see also Figure 1)

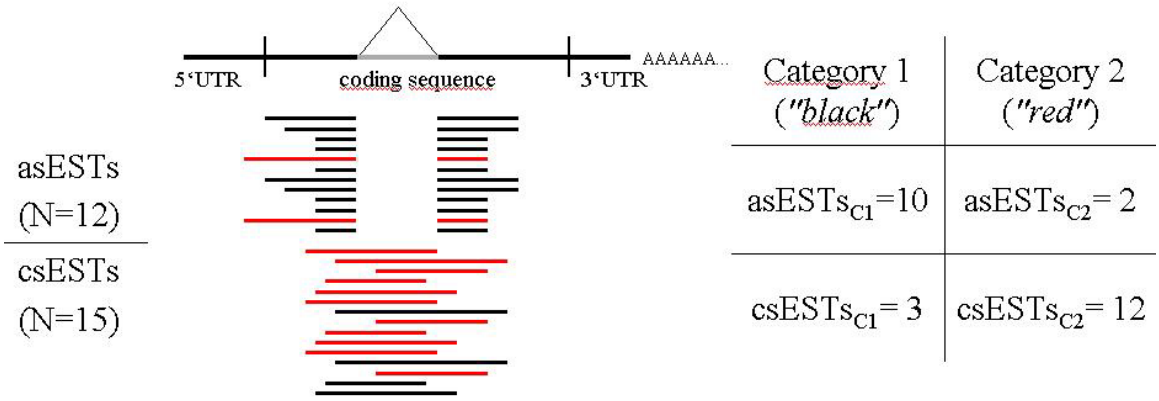

Fig.1: The strategy to categorise ESTs. In the left part, all matching ESTs are shown. The upper 12 ones are those supporting the alternative splice site (in this case: a skip of nucleotides, marked in grey). The 15 ESTs in the lower part match perfectly with the coding sequence and are defined as normally (or constitutively) spliced ESTs. Moreover, the colors red and black give the classification to category 1 and category 2, respectively. The cross table for this example is shown in the right part of this figure.

Due to the mentioned heterogeneity in the number of ESTs per library, the biased numbers of asESTs and csESTs per splice site and the possible low quality of ESTs have to be considered. Therefore, the distance measure **asDcs** (**D**istance from **as**ESTs to **cs**ESTs) was developed.

2.2.1 Scaling EST counts via Hill functions

For scaling the EST counts to the range 0 to 1, we introduce constants $\alpha,\beta>0$ and the following modified Hill function h defined by

$$h(k) = \frac{k^{\beta}}{\alpha^{\beta} + k^{\beta}} \tag{1}$$

for each $k \in \mathrm{N}$.

α is the inflection point and the following holds:

- if $k < \alpha$, then $h(k) < 0.5$ is very close to 0.
- if $k = \alpha$, then $h(k) = 0.5$.
- if $k > \alpha$, then $h(k) > 0.5$ is very close to 1.

The Hill coefficient β influences the slope of $h(k)$. The smaller β, the smoother the transition from 0 to 1.

2.2.2 The distance measure

We consider the scaled EST counts as two points in a coordinate system in $(\mathrm{P} \cap [0,1])^2$. The x-axes is used for the scaled EST counts falling in category $c1$ and the y-axes for the scaled EST counts falling in category $c2$. That is,

$p_{as} := (h(ec(as,c1)),h(ec(as,c2)))$ and

$p_{cs} := (h(ec(cs,c1)),h(ec(cs,c2)))$

The distance (p,p') of any two pints $p=(x,y)$ and $p'=(x',y')$ in this coordinate system is

$$\delta(p,p') = \sqrt{|x-x'| + |y-y'|} \tag{2}$$

We are interested in

$$\delta_{\min}(p) := \min\{\delta(p,(x',x')) \mid x' \in \text{☼} \cap [0,1]\}.$$

The points (x',x') represent the even division of the asESTs respectively csESTs into the categories $c1$ and $c2$. Let (x',x') be the point such that $\delta_{\min}(p)=\delta(p,(x',x'))$. Then, $|x-x'|=|y-x'|$ that results in

$$x' = \frac{x+y}{2} \tag{3}$$

Considering (2) and (3), $\delta_{\min}(p)$ can be observed as:

$$\delta_{\min}(p) = \frac{1}{2}\sqrt{2}|x-y| \tag{4}$$

We finally obtain the following distance measure:

$$asDcs(p_{as}, p_{cs}) := \delta_{\min}(p_{as}) + \delta_{\min}(p_{as})$$

$$asDcs(p_{as}, p_{cs}) = \begin{cases} \frac{1}{2}\sqrt{2}\left|h(ec(as,c1)) - h(ec(as,c2))\right| + \frac{1}{2}\sqrt{2}\left|h(ec(cs,c1)) - h(ec(cs,c2))\right| \\ \quad \text{if } h(ec(as,c1)) \le h(ec(as,c2)) \text{ and } h(ec(cs,c1)) \ge h(ec(cs,c2)) \text{ or} \\ \quad\quad h(ec(as,c1)) \ge h(ec(as,c2)) \text{ and } h(ec(cs,c1)) \le h(ec(cs,c2)) \\ 0 \quad \text{otherwise} \end{cases} \tag{5}$$

Hence, the maximal distance for asDcs is $\sqrt{2}$.

As already mentioned, we are interested in specifically alternatively spliced forms, e.g. those splice sites where the majority of alternatively spliced ESTs falls in one category whereas the majority of normally spliced ESTs falls in the second category (inverse ratio of ESTs). The inverse ratio has to be checked before calculating *asDcs* (see the constraint in (5)).

3 Results

The latest release of EASED (March 2005) contains 16,312 alternatively spliced human genes with 251,059 distinct splice sites. The average number of adult ESTs per alternative splice site is 30.45, whereas 13.58 embryonic ESTs support one alternative splice site on average. 79,560 of these splice sites are supported by at least one EST from adult tissue and by at least one embryonic EST.

An inverse ratio of adult to embryo ESTs for alternatively spliced ESTs compared to the ratio for normally spliced ESTs was found for 17,326 splice sites. The ranking of these potential specifically spliced forms depend on the selected input parameters α and β. To vary these parameters, the site http://cardigan.zbh.uni-hamburg.de/asDcs can be used. Additionally, it is possible to restrict the analysis only to splice sites with a certain minimal splice event length. The resulting list of potentially embryo specific alternatively spliced candidates is sorted by the *asDcs*-value (5).

The effect of several combinations of α and β for *asDcs* was further analysed. As espected, the Hill coefficient β defines the strictness of the scaling of the true EST values. Therefore, β influences the *asDcs*-values and gives the slope of the resulting graph. Figure 2 plots *asDcs* for a fixed value of α and for β in the range 1 to 10.

Additionally, α influences the ranking of the splice candidates. Table 1 shows the top ranked splice sites for 12 values of α. The upper part contains the best results for each of these 12 values of α. For 6,818 splice sites, the splice event comprises a nucleotide stretch of 30 bp or more. The best ranked splice sites of those splice sites are listed in the lower part of Tab.1.

4 Conclusions

The aim of the analysis presented here was to show how embryo specific alternatively spliced forms could be identified under the assumption that ESTs used for splice site prediction represent the occurrence of stage specific isoforms. The proposed *asDcs*-value was introduced to handle the unequal distribution of ESTs per splice site on one hand, and the possible uncertainties due to the relatively low quality of EST data on the other hand. The sigmoid shape of the Hill curve was chosen because of its asymptotic character that addresses

these both points: For large EST counts (above the inflection point α) *h(k)* tend to 1 regardless the true number of ESTs. Therefore, overrepresented ESTs influence the distance measure in the same manner as moderately represented ESTs. On the other hand, for very small EST counts (below the inflection point α) *h(k)* scales the number of ESTs to 0. Hence, very small ESTs counts (that are possibly falsely annotated or falsely aligned) are more or less disregarded. To decide what minimal number of ESTs should be disregarded lies in the hand of the user be varying the parameter α. α acts as a kind of threshold. This effect of the chosen parameter α is exemplified at Table 1: For a low value of α, the splice sites with a clear unequal EST ratio are the top of the list (e.g. Q96CP1 is on rank 1 for α=1 or α=2). If it is important to obtain splice sites with many supporting ESTs, one chooses a larger value for α and the top of the list shows splice sites with a less clear EST ratio but with large EST numbers. HNRPA1, for instance, do not have such a clear unequal EST ratio as Q96CP1 (therefore, it was only found on place 17,279 for α=1), but the EST numbers for HNRPA1 exceeds that of Q96CP1 clearly.

The slope of the curve for the Hill function can be interpreted as the strictness of the discrimination between EST counts that can be neglected and EST counts that are assumed as maximal. This slope is influenced by the Hill coefficient β, which results in a smoother transition if it becomes smaller. Since the *asDcs* value also depends on β, this can be observed for the distribution of the values of *asDcs*, too (cf. Fig. 2).

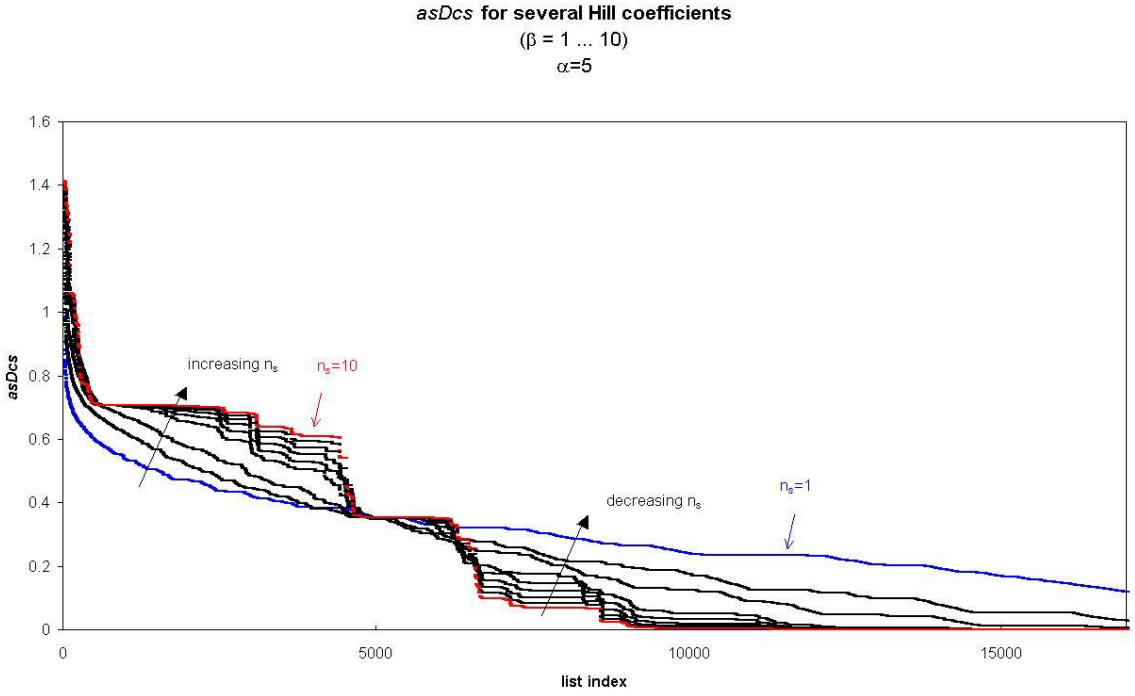

Fig.2: Illustration of the calculated distance measures *asDcs* for several Hill coefficients β (see arrows) for all 17,326 splice sites with an inverse ratio of adult to embryo ESTs for alternatively spliced ESTs compared to the ratio for normally spliced ESTs. *asDcs* was plotted over the list index of the sorted splice site lists. Every list for each Hill coefficient was sorted separately.

The second step was to calculate the distance measure *asDcs*. We calculate the distance of the pairs for both categories that are to be compared. Because of the defined interval for *asDcs*, the calculated measure distances can be easily compared.

The values of α and β depends on what is important for a certain analysis. That is thy, a web interface was developed that allows to determine the values α and β before computing the desired ranking of splice sites.

A literature search for the genes listed in Table 1 revealed the plausibility of the presented analysis. The transcription factor P65 (Q96CP1) plays a pivotal role in morphogenesis and embryonic development in mice [Nishikimi et al. 1999]. Drebrin (DBN1) is a development-associated brain protein from rat embryos [Ishikawa et al. 1994]. ATF4 was shown to be critical for normal cellular proliferation, especially for the high-level proliferation required during fetal-liver hematopoiesis [Masuoka & Townes 2002]. Finally, the AHSG/fetuin gene may have a role in differentiation since it is expressed in mouse limb buds and brain only at certain stages during development [Yang et al. 1992]. These findings point out that all the found genes are highly regulated during the embryogenesis. Further analyses are needed to clarify the real effect of alternative splicing of these genes for the embryonal development.

This brief study demonstrates the capability of the approach presented here. The distance measure is easy to calculate and to understand on one hand, but very effective on the other hand. This approach can further be extended for the identification of candidate genes for other kinds of specificity, as e.g. cancer or tissue specificity.

Table 1: Summary of the results for several values of α and a fixed Hill coefficient β=5. Only, the top ranked splice sites are listed. The first three colomns cover the gene information, the following four colomns contain the splice profiles with the (unscaled) numbers of alternatively and constitutively spliced ESTs, distinguishing between adult and embryonic origin. The last 12 columns comprises the position of the splice site for the denoted values of α. The upper part covers all splice sites regardless of their length. In the lower part, the top ranked transcripts with a splice event length = 30 bp are listed.

Gene		Swiss-Prot	alternatively spliced ESTs		constitutively spliced ESTs		splice event length	Ranking within the list for several values of α											
Name	Description	id	adult	embryo	adult	embryo	bp	α =1	α =2	α =3	α =4	α =5	α =6	α =7	α =8	α =9	α =10	α =50	E=100
Q96CP1	Transcription factor p65	Q04206	0	7	9	0	2	**1**	**1**	3	10	20	32	92	773	2729	3420	7041	7164
DBN1	Drebrin	Q16643	0	11	13	1	3	596	9	**1**	**1**	**1**	2	3	7	19	37	5267	5458
Q9H836	Integrase interactor 1 protein	Q12824	2	14	12	1	27	9234	531	30	3	2	**1**	2	2	8	19	4794	5012
HPR	Haptoglobin-related protein precursor	P00739	4	28	16	1	3	11355	2744	690	145	27	8	**1**	**1**	**1**	**1**	2448	2857
ATF4	Activating transcription factor 4	P18848	26	71	123	12	2	17251	17251	17251	14010	12164	10536	8898	7357	6175	4814	**1**	181
HNRPA1	Heterogeneous nuclear ribonucleoprotein A1	P09651	58	183	128	22	5	17279	17279	17279	17279	16384	14711	13015	12716	11082	9805	4	**1**
PIK4CA	Phosphatidylinositol 4-kinase alpha	P42356	0	8	7	0	828	**2**	**2**	4	11	24	50	144	2006	3520	4136	7503	7607
AHSG	Fetuin-A	P02765	8	0	1	87	55	732	12	**2**	**2**	7	11	18	33	59	68	230	704
HPR	Haptoglobin-related protein precursor	P00739	1	25	10	2	174	9311	544	31	4	**3**	**3**	**4**	8	16	28	2863	3206
AHSG	Fetuin-A	P02765	17	0	6	135	55	3690	2170	1111	524	159	49	16	**5**	**3**	**3**	38	260
na*	*ENSG00000194833***	na*	0	1	404	2	3166	9022	6912	3364	1675	999	669	470	326	264	222	**10**	9
COL17A1	Bullous pemphigoid antigen 2	Q9UMD9	1	2	569	7	610	16088	9491	9578	10381	8570	5800	4139	3062	2259	1627	11	**8**

* na=not available ** only the ENSEMBLE id was available

5 Acknowledgement

I wish to acknowledge Alexander Herrmann (Max-Delbrück-Centrum für Molekulare Medizin Berlin-Buch, Germany) for technical realisation of the EASED project and Stefan Kurtz (ZBH, Universität Hamburg) for proof-reading.

6 References

[1] A. A. Mironov, J. W. Fickett, M. S. Gelfand. Frequent alternative splicing of human genes. Genome Res, 9(12):1288-1293, 1999.

[2] D. Brett, J. Hanke, G. Lehmann, S. Haase, S. Delbruck, S. Krueger, J. Reich, P. Bork. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. FEBS Lett., 474(1):83-86, 2000.

[3] Z. Kan, E. C. Rouchka, W. R. Gish, D. J. States. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res., 11(5):889-900, 2001.

[4] C. Lee, L. Atanelov, B. Modrek and Y. Xing. ASAP: the alternative splicing annotation project. Nucleic Acids Research, 31(1):101–105, 2003.

[5] B. Modrek, A. Resch, C. Grasso, C. Lee. Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res., 29(13):2850-2859, 2001.

[6] J. M. Johnson, J. Castle, P. Garrett-Engele, Z. Kan, P. M. Loerch, C. D. Armour, R. Santos, E. E. Schadt, R. Stoughton, D. D. Shoemaker. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science, 302(5653):2141-2144, 2003.

[7] T. A. Thanaraj, S. Stamm, F. Clark, J. J. Riethoven, V. Le Texier and J. Muilu. ASD: the Alternative Splicing Database. Nucleic Acids Research, 32(Database issue):D64-D69, 2004.

[8] I. Dralyuk., M. Brudno, M. S. Gelfand., M. Zorn and I. Dubchak. ASDB: database of alternatively spliced genes. Nucleic Acids Research, 28(1):296-297, 2000.

[9] N. Kim, Y. Shin and S. Lee. ECgene: genome annotation for alternative splicing. Genome Research, 15(4):566-576, 2005.

[10] M. Burset, I. A. Seledtsov and V. V. Solovyev. SpliceDB: database of canonical and non-canonical mammalian splice sites. Nucleic Acids Research, 29(1):255-259, 2001.

[11] A. Krause, S. A. Haas, E. Coward and M. Vingron. SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. Nucleic Acids Research, 30(1):299-300, 2002.

[12] H. Pospisil, A. Herrmann, R. H. Bortfeldt and J. G. Reich. EASED: Extended Alternatively Spliced EST Database. Nucleic Acids Research, 32(Database issue):D70-D74,2004.

[13] E. Birney, D. Andrews, P. Bevan, M. Caccamo, G. Cameron, Y. Chen, L. Clarke, G. Coates, T. Cox, J. Cuff, V. Curwen, T. Cutts, T. Down, R. Durbin, E. Eyras, X.M. Fernandez-Suarez, P. Gane, B. Gibbins, J. Gilbert, M. Hammond, H. Hotz, V. Iyer, A. Kahari, K. Jekosch, A. Kasprzyk, D. Keefe, S. Keenan, H. Lehvaslaiho, G. McVicker,

C. Melsopp, P. Meidl, E. Mongin, R. Pettett, S. Potter, G. Proctor, M. Rae, S. Searle, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, A. Ureta-Vidal, C. Woodwark, M. Clamp and T. Hubbard. Ensembl. 2004. Nucleic Acids Research, 32(Database issue):D468-D470, 2004.

[14] W. Gish (1996-2004) [http://blast.wustl.edu]

[15] L. Florea, G. Hartzell, Z. Zhang, G. M. Rubin and W. Miller. A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Research, 8(9):967-974, 1998.

[16] J.A. Bedell I. Korf and W. Gish. MaskerAid: a performance enhancement to RepeatMasker. Bioinformatics, 16:1040-1041, 2000.

[17] A. Nishikimi, J. Mukai and M. Yamada. Nuclear translocation of nuclear factor kappa B in early 1-cell mouse embryos. Biol Reprod, 60(6):1536-1541, 1999.

[18] R. Ishikawa, K. Hayashi, T. Shirao, Y. Xue, T. Takagi, Y. Sasaki and K. Kohama. Drebrin, a development-associated brain protein from rat embryo, causes the dissociation of tropomyosin from actin filaments. J Biol Chem, 269(47):29928-29933, 1994.

[19] H.C. Masuoka and T.M. Townes. Targeted disruption of the activating transcription factor 4 gene results in severe fetal anemia in mice. Blood, 99(3):736-745, 2002.

[20] F. Yang, Z.L. Chen, J.M. Bergeron, R.L. Cupples and W.E. Friedrichs. Human alpha 2-HS-glycoprotein/bovine fetuin homologue in mice: identification and developmental regulation of the gene. Biochim Biophys Acta, 1130(2):149-156, 1992.