Research

PRESTA: associating promoter sequences with information on gene expression

Václav Mach

Address: Institute of Entomology, Czech Academy of Sciences, Branisovská 31, 370 05 Ceské Budejovice, Czech Republic.
E-mail: mach@entu.cas.cz

Published: 21 August 2002

Genome ***Biology*** 2002, **3(9)**:research0050.1–0050.7

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/9/research/0050

© 2002 Mach, licensee BioMed Central Ltd
(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 8 January 2002
Revised: 12 April 2002
Accepted: 24 June 2002

Abstract

Background: Large sets of well-characterized promoter sequences are required to facilitate the understanding of promoter architecture. The major sequence databases are a prospective source of upstream regulatory regions, but suffer from inaccurate annotation. The software tool PRESTA (PRomoter EST Association) presented in this study is designed for efficient recovery of characterized and partially verified promoters from GenBank and EMBL libraries.

Results: The PRESTA algorithm examines the putative GenBank/EMBL promoters and automatically removes most of the poorly annotated entries. The remaining records are connected to expressed sequence tags (ESTs) through a high-stringency BLAST search. The frequency and source of recovered ESTs provide an estimate of the activity and expression pattern of the promoter, and the ESTs' 5′ ends assist in transcription start-site verification. The PRESTA database provides easy access to non-redundant upstream regulatory regions recently extracted by the PRESTA algorithm. The current size of this resource is 552 human and 241 mouse promoters. Surprisingly, no overlap between the PRESTA database and the Eukaryotic Promoter Database (EPD) was detected by sequence comparison.

Conclusions: The PRESTA algorithm demonstrates the principle of promoter verification by mapping EST 5′ ends. The publicly available PRESTA database collects hundreds of characterized and partially verified promoter sequences and is complementary to other promoter databases.

Background

Construction of large sets of well-characterized promoter sequences is driven by a general expectation that detailed comparative studies exploiting large groups of promoters will eventually lead to an understanding of promoter architecture. The two prominent currently available resources of upstream promoter sequences are the Transcription Regulatory Regions Database (TRRD) [1] and the Eukaryotic Promoter Database (EPD) [2]. They rely on published data and frequently neglect experimentally characterized but otherwise unpublished promoters submitted to GenBank [3] and EMBL [4] libraries. This suggests the idea of exploiting GenBank and EMBL as an alternative source of annotated upstream regulatory sequences. Unfortunately, any organism-wide GenBank/EMBL query intended to retrieve promoter sequences will recover records containing data of very variable quality. Removal of inadequate or poorly annotated entries requires considerable manual effort and can be greatly facilitated by a text-analyzing algorithm. In addition, a set of putative upstream promoter sequences is, by itself, of a limited use, as we are also interested in the expression pattern and total activity of individual promoters. Finally, an independent verification of the transcription start site is highly desirable.

Connecting the promoter sequences to the Expressed Sequence Tag (EST) database [5,6], which contains millions of end-sequenced cDNA clones derived from thousands of different cDNA libraries, provides a solution to the two last-mentioned problems. The tissue source of each cDNA library is known and can be traced from an individual EST clone. Thus, by identifying the set of ESTs corresponding to the downstream transcribed region of a particular promoter, we recover substantial information about the promoter's activity. Moreover, eventual mapping of an EST 5´ end onto the vicinity of the transcription start site increases the reliability of the +1 nucleotide placement.

The PRESTA (PRomoter EST Association) algorithm introduced in this study facilitates extraction of promoter sequences located immediately upstream of the transcription start site from both GenBank [3] and EMBL [4]. The program automatically removes most of the unwanted records recovered by a GenBank/EMBL query and presents the retained entries in a structured form. Promoters are stored together with a short stretch of the immediately downstream transcribed sequence, which serves for promoter association with matching ESTs through the use of a high-stringency BLAST search [7]. The frequency and source of these ESTs provide an estimate of the promoters' activity and expression pattern, whereas their 5´ ends assist in transcription start site verification. The non-redundant GenBank/EMBL human and mouse promoters recently extracted by PRESTA are stored in the PRESTA database [8] and are interactively accessible via the web.

Results and discussion

Promoter retrieval

A common weakness of the major sequence libraries is the diverse quality of their entries. This became apparent when the GenBank database was screened for putative human and mouse promoter sequences using the keywords '5´UTR', 'prim_transcript', 'tata_signal' and 'promoter'. On inspecting the results, only a small fraction out of the 2,300 human and 1,500 mouse entries obtained was found to contain annotated promoter sequence of sufficient length. Most records were eliminated, but some others could be converted into more than one sequence and thus the final number of retained upstream promoter regions was 470 and 280 human and mouse entries, respectively. All manipulations were carried out using the PRESTA data presentation environment, and the average time required to process an entry was approximately 10 seconds.

The original data were subsequently reanalyzed using a pre-filtering text-analysis tool that allows PRESTA to automatically remove apparently irrelevant entries (see Materials and methods). This compressed the size of the initial pool into 482 (human) and 311 (mouse) GenBank records. Upon inspection, 85 and 62 entries, respectively, were removed, and the remaining items were converted into a set of promoters which was identical to the sample obtained when pre-filtering was not used. This was probably a coincidence, as text analysis is a complex problem and the algorithm must make occasional errors. Even so, a test at this scale would undoubtedly reveal any serious problem in the design of the text-analysis module and the pre-filtering tool was used for all subsequent analyses.

The promoters found in this initial test were not used further. Instead, the database search was repeated using both GenBank and EMBL libraries and a more complex query (Table 1). This recovered 12,000 human and 5,500 mouse entries, a number that would be hardly manageable without automation. The default pre-filtering tool setting prevented acceptance of promoters associated with features described by words such as 'putative', 'not_experimental', and so on. Similarly, when the pre-filtered pools were refined, data originating from human and mouse genome sequencing projects were avoided, unless it was apparent that the annotation is based on either an experimental observation or high similarity to an experimentally annotated gene. The recovered promoters were transiently stored in the PRESTA internal 'tag' format (Table 1).

As explained above, an important function of the PRESTA algorithm is the selection of reliably annotated GenBank/EMBL entries. This aim is shared by RefSeq [9], a highly reliable human-curated sequence database mostly derived from GenBank. Although using RefSeq instead of GenBank would undoubtedly streamline PRESTA operations, this is not currently feasible because of the low number of full-length RefSeq cDNAs (data not shown).

Connecting promoters and ESTs

PRESTA connects promoters to ESTs through the use of a high-stringency BLAST search (see Materials and methods). The recovered ESTs serve two purposes - verifying the promoter structure and providing expression data. The number of promoters successfully associated with ESTs was 271 (human, GenBank), 571 (human, EMBL), 208 (mouse, GenBank) and 150 (mouse, EMBL). Removal of duplicates reduced the total number of available EST-associated promoters to 552 human and 241 mouse entries (Table 1). The average length of an upstream promoter sequence was 750 bases.

Not all the recovered promoters are equally reliable. For example, 281 human and 124 mouse promoters are confirmed by at least two EST 5´ ends mapping within the -5 to +30 region respective to the transcription start site (Table 1 and [8]). Confirmation of an already annotated promoter by an EST is a conservative requirement and this subset of the PRESTA database [8] is unlikely to contain a significant fraction of mismapped sequences. The size of this resource is roughly comparable to the corresponding EPD figures ([2] and see also Table 1).

Table 1

Data imported by PRESTA

	Human	Mouse
EPD		
Total entries	276	200
Imported by PRESTA*	214	167
tcg total†‡	139	109
Present in GenBank/EMBL§	0	0
Weak promoters¶	4	3
Confirmed by one EST#	99	64
Confirmed by two ESTs#	82	56
GenBank		
Total entries¥	5,870	3,289
After pre-filter	570	307
tag‡	484	313
tcg total†‡	291	208
tcg non-redundant	241	192
Not found in EMBL**	128	96
EMBL		
Total entries¥	6,314	2,251
After pre-filter	1051	274
tag created‡	820	222
tcg total†‡	571	150
tcg non-redundant‡	425	145
Not found in GenBank**	312	49
GenBank + EMBL		
tcg non-redundant	553	241
Present in EPD§	0	0
Possibly misannotated¶	30	16
Confirmed by one EST#	326	153
Confirmed by two ESTs#	281	124

EPD promoters are shown for comparison. Some EPD entries did not meet the PRESTA limit on downstream sequence length. †Fraction of promoters successfully associated with ESTs. ‡Both 'tag' and 'tcg' are internal PRESTA formats, 'tag' stores the promoter sequences, 'tcg' adds information about matching ESTs. §No overlap between the GenBank/EMBL non-redundant set and PRESTA-imported EPD entries was found using pairwise SEQALN alignment of immediately downstream transcribed sequences. This is not an error: EMBL sequences linked from EPD were correctly dissected, as some of them are homologous to dozens of 5′ EST ends. Even more surprisingly, there is no apparent overlap between PRESTA and the full human subdivisions of EPD. An EPD entry directly stores a 49-base-pair stretch of the immediately upstream region. The full set of these stretches was downloaded by a simple web agent and compared to an analogous set of PRESTA sequences using SEQALN. ¶There are no ESTs confirming the transcription start site and at least two 5′ EST ends are longer then expected. #The 5′ end of at least one (or two) matching ESTs maps to the -5 to +30 region relative to the transcription start site. In addition, the ratio of positively mapping to overshooting 5′ ends is larger than 1:3. The current PRESTA version neglects the possibility that the library was amplified and that two or more ESTs actually originate from the same cDNA clone. ¥A sample query: (([genbank-Division:rod] & (([genbank-Organism:Mus] & [genbank-Organism:musculus*]) | [genbank-Organism:Mus musculus*])) & (((((([genbank-FtKey:5\′utr] | [genbank-FtKey:precursor_rna]) | [genbank-FtKey:prim_transcript]) | [genbank-FtKey:promoter]) | [genbank-FtKey:tata_signal]) > parent)). **Not recovered by an equivalent query. This reflects different feature annotation rather then incomplete synchronization between the two major sequence databases.

Nevertheless, a significant fraction of PRESTA promoters was not confirmed by 5′ EST mapping. Therefore, an attempt was made to exploit the increased accuracy and sensitivity of the new generation of promoter-predicting tools [10-15]. The programs FirstEF [13], Promoter Inspector [14] and EpoNINE [15] were used to analyze the promoter pools corresponding to subsets of either PRESTA or EPD databases. Unfortunately, the results are not conclusive (see predictions.html in Additional data files) as the EPD analysis set shows that genuine promoter sequences are frequently missed by the prediction algorithms. Thus, it is not clear what is the real degree of certainty added by positive prediction by one or more of the promoter-predicting tools.

It is important to note that the number of promoters recovered by PRESTA will increase. Not only will additional experimentally characterized upstream sequences be submitted to GenBank/EMBL and potentially be recovered, but the total number and quality of available ESTs will increase with the progress of large-scale sequencing projects. Consequently, the proportion of promoter sequences associated with expression information will also increase. In future, it should become possible to extend the principle used by PRESTA into promoter annotation *in silico*, solely on the basis of mapping EST 5′ ends. This concept is similar to the promoter-prediction method recently used by Liu and States [16]. Instead of simultaneously mapping multiple EST ends, however, the authors narrow the putative promoter region by extending the 5′ end of a single EST with the aid of gene-modeling software. There is an obvious difference between the current version of PRESTA and any future promoter-prediction tool based on the 5′ EST mapping: PRESTA accesses GenBank and EMBL promoters which were annotated on the basis of experimental data.

Expression pattern

The large-scale resources of gene-expression information include SAGE (serial analysis of gene expression) libraries [17], data obtained using microarray techniques [18,19] and information embedded in EST [5,6] libraries. The latter resource is exploited by PRESTA.

To evaluate the quality of the PRESTA-generated expression pattern, EPD mouse promoters were imported into the PRESTA environment (Table 1) and their annotation was compared with the tissue source of matching ESTs. This comparison did not include promoters of genes abundantly expressed in lymphocytes, as EST libraries are frequently contaminated by blood cells. These promoters were excluded on the basis of functional description of the matching ESTs, conveniently viewed within the PRESTA environment.

A significant proportion of the EPD promoters correspond to housekeeping genes. Identifying a housekeeping or otherwise ubiquitous activity pattern *in silico* is a challenging task. Although PRESTA cannot positively confirm that a

gene is ubiquitously expressed, an exceptional diversity in the tissue origin of matching ESTs is a strong indication of the housekeeping nature of the promoter. In the case of the EPD mouse housekeeping promoters, more than a third of them matched ESTs derived from at least 16 different libraries (data not shown). Thus, PRESTA's ability to recognize a general expression pattern is at present limited, but is promising in view of the expected increase in the number of available EST clones. Housekeeping genes are not expressed uniformly [20,21]. The semi-quantitative data provided by PRESTA show large differences in the total level of activity among individual housekeeping promoters, varying from 1 to 100 matching ESTs (data not shown).

The degree of correlation between PRESTA and EPD is particularly interesting for promoters displaying a restricted expression pattern, as it provides an estimation of the reliability of PRESTA data. Table 2 lists the unambiguously expressed EPD mouse promoters, together with EPD and PRESTA expression information. With few exceptions, the agreement is remarkably good.

The discussion so far has revolved around the tissue-specific pattern of activity of the promoter. But, as shown in Table 2, PRESTA also reports the total numbers of matching ESTs. These figures cannot, however, be treated as a truly quantitative estimate of the promoter's strength. For example, some promoters may appear more active than others simply owing to the large number of ESTs available for a specific tissue. Furthermore, the BLAST search relies on the immediately downstream transcribed sequence and will therefore preferentially detect ESTs derived from random-primed or full-length cDNA libraries as well as ESTs corresponding to short transcripts cloned in oligo(dT)-primed libraries. In this way, a promoter may appear more intensively expressed in a certain tissue just because of the availability of a

Table 2

Correlation of EPD and PRESTA promoter expression patterns

EPD entry	EPD ('DO' expression field)	PRESTA (tissue or organ, number of ESTs)
MM_ALBU	Liver	Liver, liver tumor 93 , other 8
MM_AMYA	Liver > salivary gland, pancreas	Liver 3
MM_AMYPA	Pancreas strong	Pancreas 242, other 8
MM_ANF	Cardiac atrium	Heart 9, pooled 7, lung 1, kidney 1
MM_BGLR	Kidney	Kidney 4, mammary 9, lung 2, colon 1, other 10
MM_CA13	Fibroblasts	Various
MM_COLI	Pituary	Pituary 1
MM_CRA2	Lens	Retina, lens 5, embryo head 2, total fetus 11
MM_CRGF	Differentiating lens	Retina 4, fetus 24
MM_FABI	Gut	Intestine 1, liver 1
MM_FRIH	Heart > liver, spleen	Many: liver 34, kidney 31, myotubes 33, heart 2, pooled and other 150
MM_GFAP	CNS, astrocytes	Hypothalamus 1, corpora 1, other 1
MM_HSP1	Spermatids	Testis 147, unknown 13
MM_INS1	Pancreas strong	Pancreas 114, small intestine 1
MM_KLK1	Salivary gland, pancreas, spleen, testis	Kidney 1
MM_MBP	Oligodendrocytes	Brain 80, medulla 2, other 9
MM_MDR1_B	Adrenal gland, kidney, placenta	Hypothalamus 1, intestine 1, pooled 3
MM_MOS1	Testis	Unknown 1
MM_MYG	Skeletal muscle	Diaphragm 18, myotubes 4, heart 2, mammary 5, pooled and other 8
MM_MYPR	Oligodendrocytes	Brain 21, pooled and unknown 10
MM_OTC	Liver, intestine	Liver 17, colon 3, unknown 3
MM_PSP	Parotid gland	Salivary gland 24, skin 2
MM_RENS	Kidney, submaxillary gland	Salivary gland 50
MM_STP2	Spermatids	Testis 29, unknown 2
MM_TRIC	Heart	Heart 10, B cells 1, pooled 7
MM_TTHY	Liver, brain, kidney	Liver 78, brain 2, kidney 6 , lung 28, colon 9, diaphragm 12, pooled and other 78
MM_PR73	Mammary tumor	Mammary (normal) 27, lung 3, unknown 4

Includes EPD promoters from Table 1 displaying at least a partly restricted expression pattern. In some cases, the PRESTA data are more complete: ferritin heavy chain (MM_FRIH) is expressed in multiple organs [33], P-glycoprotein (MM_MDR1B) was found in the blood-brain barrier [34] and myoglobin (MM_MYG) is expressed in heart muscle [35].

full-length library. An additional complication arises from the fact that many EST libraries were subtracted, so the frequency of clones derived from rare and abundant transcripts will be higher and lower, respectively, than the actual frequency of the corresponding mRNAs.

ESTs were traditionally used for comparative evolutionary studies and for gene hunting [22]. Recently, the true potential of this resource is becoming widely recognized, and ESTs are being used for gene annotation [23-25]. The inhomogeneous nature of this resource is partly compensated for by its enormous size (currently 4 million human and more than 2 million mouse ESTs) and PRESTA data should therefore be treated as semi-quantitative.

PRESTA is not the only utility using EST expression information, as a similar principle is exploited by tools such as CGAP Gene Finder [26], DDD [27] or even advanced SRS searches [28]. These utilities are not, however, linked to gene promoter sequences and therefore not readily useful for large-scale studies of promoter expression. Such a link beween structural and functional information is provided by PRESTA and EPDEX [2]. As detailed in the legend of Table 1, the PRESTA database and EPD/EPDEX are complementary resources exploiting different promoter pools.

Materials and methods

Algorithm overview

PRESTA is available in two forms: a Windows program available for download from [8] and a searchable online database available at the same address. To avoid confusion, the database is always referred as the PRESTA database. PRESTA is implemented in Visual Basic 6.0 and serves for promoter retrieval, transcription start site verification and integration of promoter sequences with expression data. The algorithm requires Microsoft Internet Explorer, preferably version 5.0 or higher. Additional third-party utilities, including the National Center for Biotechnology Information (NCBI) BLASTCl3 client and some parts of the SEQALN algorithm [29], are installed with PRESTA. To facilitate user orientation, the PRESTA package comes with extensive help.

The PRESTA database provides online access to human and mouse promoters recently extracted by PRESTA from GenBank and EMBL libraries (Table 1). The database is hosted on MySQL platform and served by PHP scripts using the native MySQL drivers.

Source of promoter sequences

PRESTA greatly facilitates the recovery of promoter sequences from either GenBank or EMBL libraries. The GenBank/EMBL entry is presented to the user in a pre-analyzed and structured form. This makes promoter extraction a very fast process and minimizes the human involvement. Moreover, PRESTA automatically removes most of the unwanted entries present in the starting GenBank/EMBL output file. An entry will be removed if the algorithm is unable to suggest a putative transcription start site using the procedure explained later in this section.

The program can also import promoter sequences referred to in the EPD database. PRESTA was not, however, intended as an alternative EPD interface; this feature was implemented for comparative purposes and may be discontinued in future PRESTA releases.

Implementation of entry evaluation

Evaluation of the GenBank/EMBL entry is performed by context-dependent text analysis. This process starts by parsing the entry into individual features: '5´utr', 'prim_transcript', 'precursor_rna', 'promoter', 'tata_signal', 'cds', 'mRNA', 'intron', and 'exon' if this is the first exon. Introns serve for PRESTA internal use and are never presented to the user. Each feature is represented by a feature object characterized by properties such as feature type, location, qualifier, strand, and so on. The value of 'nucleotide_minus_one' property is deduced from the location of 'prim_transcript', 'precursor_rna' and '5´utr' features. The next step consists of aggregating the feature objects into collections. Rather than relying on the text order within the entry, PRESTA extensively analyzes relationships between features using location, strand and eventual gene name properties. As a result, a collection contains features associated with a single transcription start site. Finally, the algorithm refines the coordinates of the transcribed sequence located immediately downstream of the transcription start site. This is done by comparing the different feature types contained in the collection, for example the location of '5´utr' and 'cds' are frequently combined, whereas the 'intron' coordinates assist in *in silico* splicing of the 'prim_transcript'. This step includes checking for eventual uncertainty in location of some features. A collection of features is presented to the user, provided that the algorithm located the transcription start site as well as both upstream and downstream sequence of a certain minimal length (default 60 bases).

The user can alter the details of this procedure by specifying the types of features used by the program and editing feature-specific lists of 'forbidden' words. Occurrence of such words in the feature qualifier will result in rejection of the particular transcription start site.

Connecting promoters to expression information

There is no direct way of connecting upstream promoter sequence with ESTs. PRESTA therefore stores the upstream promoter area together with a short stretch of the immediately downstream transcribed sequence. This downstream sequence tag is used as a query in a high-stringency BLAST search [7] against the EST database. The exact value of the BLAST 'Expect' (*E*) parameter is not very important as

PRESTA extensively filters the BLAST result. Using the default settings, this filtering in practice removes matches with E greater than 10^{-30}. Values of E in the range 10^{-1} to 10^{-27} were used in preliminary analyses and the data presented in this study were obtained using the default value of 10^{-21}. Another variable affecting the BLAST search is the length of the sequence used as a query. Longer downstream sequences will increase the total number of recovered ESTs, but these are more likely to be derived from a downstream promoter. The default setting used in this study is 500 bp.

The description of a recovered EST includes the name of the library from which the clone originates. In this way, the activity of the promoter is described both qualitatively (the tissue source of the corresponding EST libraries) and semi-quantitatively (the total number of recovered EST clones of a certain origin).

PRESTA wraps the BLAST search, which is performed by NCBI BLASTCl3 [30] client as the background operation. Communication is indirect: PRESTA launches BLASTCl3 and, on notification, analyzes the BLAST output file. The classification of EST libraries was also originally obtained from NCBI, but was recently supplemented by the description of the human UniGene libraries [31].

File and database format

PRESTA temporarily stores data in several internal formats and the final annotated promoter sequence is stored in the 'tcg' format. The 'tcg' file contains the GenBank or EMBL annotation, the upstream promoter sequence, downstream transcribed sequence and description of matching EST clones. This format is convenient for handling the promoter-containing entries inside the PRESTA environment and is also supported by the promoter-analysis tool ELSEA [32]. Alternatively, PRESTA promoter sequences can be exported in the multiple FASTA format or most of the information can be saved in the form of 'database-ready' flat files. The flat files containing human and mouse entries recently extracted from GenBank and EMBL were imported into the public PRESTA database [8].

Promoter structure verification

In many instances, PRESTA independently confirms the GenBank/EMBL transcription start site annotation. This is done by comparing the +1 nucleotide position and the 5′ ends of recovered ESTs. Generally, 5′ ends can provide either positive or negative evidence. Positive evidence is that an EST 5′ end maps close to the transcription start. Such an observation increases the reliability of the transcription start site annotation. Conversely, an EST 5′ end extending further upstream is considered as negative evidence. Negative evidence does not necessarily indicate that the promoter's transcription start site is invalid. Frequently, some ESTs end near the transcription start site whereas others are longer. This usually suggests the existence of an alternative upstream transcription start site. On the other hand, some transcription start sites are not confirmed by any EST whereas one or more clones extend to the upstream area. This means that these promoters are either mismapped or that a strong transcription start site is located further upstream. Additional details are evident from illustrations included in the algorithm Setup.

Download of the full EST sequences is performed in the background using Microsoft Internet Explorer functions.

Sequence comparison

Although most of the sequence comparison is done outside PRESTA by the BLAST algorithm, PRESTA occasionally compares sequences for the purpose of duplicate removal and 5′ EST end-mapping. This is achieved by local similarity sequence comparison performed by the SEQALN algorithm [29].

The SEQALN code was directly incorporated into PRESTA in the form of C++ dll. Some SEQALN functions were removed, as well as most of the original interface, which was replaced by connectivity to Visual Basic. Neither of these changes directly affected the SEQALN local similarity engine. Functionality of the modified algorithm was verified using the public SEQALN server. The newly added functions interfacing SEQALN with applications written in Visual Basic are available to other developers, as detailed in PRESTA help.

Security

PRESTA downloads data from various servers over the Internet. The BLAST search is managed by the NCBI server through the NCBI BLASTCl3 client. On the other hand, PRESTA controls downloading of the EPD entries, EMBL records pointed from EPD and full EST sequences required for the 5′-end mapping. To prevent the possibility of server abuse, PRESTA's ability to query a server is strictly limited to a maximum of a single entry in 10 sec. Thus the burden imposed by the program is probably lower than the load resulting from a single user contacting the server manually. Because of this limitation, download operations are slow and generally performed in the background.

Additional data files

A file (predictions.html) containing additional verification of PRESTA promoters is available with the online version of this paper.

Acknowledgements

PRESTA incorporates BLASTCl3 client and other resources developed or maintained by NCBI as well as parts of SEQALN algorithm developed by P.M. Hardy and M.S. Waterman. I am grateful to Masako Asahina-Jindrová, Ivo Šauman and other colleagues for their help and useful suggestions. This work was partly supported by grants 204/99/1336 from the Grant Agency of the Czech Republic and K5052113 from the Grant Agency of the Czech Academy of Sciences.

References

1. Kolchanov NA, Ignatieva EV, Ananko EA, Podkolodnaya OA, Stepanenko IL, Merkulova TI, Pozdnyakov MA, Podkolodny NL, Naumochkin AN, Romashchenko AG: **Transcription regulatory regions database, TRRD: its status in 2002.** *Nucleic Acids Res* 2002, **30:**312-317.
2. Praz V, Périer RC, Bonnard,C, Bucher P: **The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data.** *Nucleic Acids Res* 2002, **30:**322-324.
3. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2002, **30:**17-20.
4. Stoesser G, Baker W, van den Broek A, Camon E, Garcia-Pastor M, Kanz C, Kulikova T, Leinonen R, Lin Q, Lombard V, *et al.*: **The EMBL Nucleotide Sequence Database.** *Nucleic Acids Res* 2002, **30:**21-26.
5. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, *et al.*: **Complementary DNA sequencing: expressed sequence tags and the human genome project.** *Science* 1991, **252:**1651-1656.
6. Boguski MS, Lowe TM, Tolstoshev CM: **dbEST-database for "expressed sequence tags".** *Nat Genet* 1993, **4:**332-333.
7. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
8. **PRESTA database** [http://baloun.entu.cas.cz/presta]
9. Pruitt KD, Maglott DR: **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 2001, **29:**137-140.
10. Fickett JW, Hatzigeorgiou AG: **Eukaryotic promoter recognition.** *Genome Res* 1997, **7:**861-878.
11. Pedersen AG, Baldi P, Chauvin Y, Brunak S: **The biology of eukaryotic promoter prediction - a review.** *Comput Chem* 1999, **23:**191-207.
12. Hannenhalli S, Levy S: **Promoter prediction in the human genome.** *Bioinformatics* 2001, **17 (Suppl 1):**S90-S96.
13. Davuluri RV, Grosse I, Zhang MQ: **Computational identification of promoters and first exons in the human genome.** *Nat Genet* 2001, **29:**412-417.
14. Scherf M, Klingenhoff A, Werner T: **Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach.** *J Mol Biol* 2000, **297:**599-606.
15. Down TA, Tim JP, Hubbard TJP: **Computational detection and location of transcription start sites in mammalian genomic DNA.** *Genome Res* 2002, **12:**458-461.
16. Liu R, States DJ: **Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling.** *Genome Res* 2002, **12:**462-429.
17. Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, Altschul SF: **SAGEmap: a public gene expression resource.** *Genome Res* 2000, **10:**1051-1060.
18. Hegde P, Qi R, Abernathy K, Gay C, Dharap S, Gaspard R, Hughes JE, Snesrud E, Lee N, Quackenbush J: **A concise guide to cDNA microarray analysis.** *Biotechniques* 2000, **29:**548-556.
19. **GEO Accession Display Tool** [http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi]
20. Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, Grisar T, Igout A, Heinen E: **Housekeeping genes as internal standards: use and limits.** *J Biotechnol* 1999, **75:**291-295.
21. Kletzien RF, Harris PK, Foellmi LA: **Glucose-6-phosphate dehydrogenase: a "housekeeping" enzyme subject to tissue-specific regulation by hormones, nutrients, and oxidant stress.** *FASEB J* 1994, **8:**174-181.
22. Pandey A, Lewitter F: **Nucleotide sequence databases: a gold mine for biologists.** *Trends Biochem Sci* 1999, **24:**276-280.
23. Ewing RM, Kahla AB, Poirot O, Lopez F, Audic S, Claverie JM: **Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression.** *Genome Res* 1999, **9:**950-959.
24. Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J: **An optimized protocol for analysis of EST sequences.** *Nucleic Acids Res* 2000, **28:**3657-3665.
25. Muilu J, Rodriguez-Tome P, Robinson A: **Gbuilder - an application for the visualization and integration of EST cluster data.** *Genome Res* 2001, **11:**179-184.
26. **CGAP Gene Finder** [http://cgap.nci.nih.gov/Genes/GeneFinder]
27. **Digital Differential Display Guide** [http://www.ncbi.nlm.nih.gov/UniGene/info_ddd.shtml]
28. Zdobnov EM, Lopez R, Apweiler R, Etzold T: **The EBI SRS server - recent developments.** *Bioinformatics* 2002, **18:**368-373.
29. **SEQALN algorithm** [http://www-hto.usc.edu/software/seqaln/]
30. **The BLAST network client software** [http://www.ncbi.nlm.nih.gov/BLAST/blast_FAQs.html#Batch]
31. **NCBI-UniGene library browser** [http://www.ncbi.nlm.nih.gov/UniGene/]
32. **ELSEA homepage** [http://www.entu.cas.cz/mach/elsea/elsea.html]
33. Yang DC, Wang F, Elliott RL, Head JF**: Expression of transferrin receptor and ferritin H-chain mRNA are associated with clinical and histopathological prognostic indicators in breast cancer.** *Anticancer Res* 2001, **21:**541-549.
34. Rao VV, Dahlheimer JL, Bardgett ME, Snyder AZ, Finch RA, Sartorelli AC, Piwnica-Worms D: **Choroid plexus epithelial expression of MDR1 P glycoprotein and multidrug resistance-associated protein contribute to the blood-cerebrospinal-fluid drug-permeability barrier.** *Proc Natl Acad Sci USA* 1999, **96:**3900-3905.
35. Bassel-Duby R, Grohe CM, Jessen ME, Parsons WJ, Richardson JA, Chao R, Grayson J, Ring WS, Williams RS: **Sequence elements required for transcriptional activity of the human myoglobin promoter in intact myocardium.** *Circ Res* 1993, **73:**360-366.