D90–D94 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj049

HTPSELEX—a database of high-throughput SELEX libraries for transcription factor binding sites

Vidhya Jagannathan, Emmanuelle Roulet, Mauro Delorenzi and Philipp Bucher*

Swiss Institute for Experimental Cancer Research and Swiss Institute of Bioinformatics,
Ch. des Boveresses 155, CH-1066 Epalinges, Switzerland

Received August 15, 2005; Revised and Accepted October 4, 2005

ABSTRACT

HTPSELEX is a public database providing access to primary and derived data from high-throughput SELEX experiments aimed at characterizing the binding specificity of transcription factors. The resource is primarily intended to serve computational biologists interested in building models of transcription factor binding sites from large sets of binding sequences. The guiding principle is to make available all information that is relevant for this purpose. For each experiment, we try to provide accurate information about the protein material used, details of the wet lab protocol, an archive of sequencing trace files, assembled clone sequences (concatemers) and complete sets of *in vitro* selected protein-binding tags. In addition, we offer in-house derived binding sites models. HTPSELEX also offers reasonably large SELEX libraries obtained with conventional low-throughput protocols. The FTP site contains the trace archives and database flatfiles. The web server offers user-friendly interfaces for viewing individual entries and quality-controlled download of SELEX sequence libraries according to a user-defined sequencing quality threshold. HTPSELEX is available from ftp://ftp.isrec.isb-sib.ch/pub/databases/htpselex/ and http://www.isrec.isb-sib.ch/htpselex.

INTRODUCTION

SELEX (systematic evolution of ligands by exponential enrichment) is an *in vitro* protocol for isolating nucleic acid ligands to a specific protein from DNA or RNA sequences (1). The technique is frequently used for the purpose of characterizing the binding specificity of a transcription factor. In such an experiment, SELEX yields a library of double-stranded DNA molecules binding to the protein, which can then be used to generate a computational model, e.g. a position-specific scoring or weight matrix (2) that serves to predict binding sites in regulatory DNA sequences. Alternatively, the technique can also be used to select a single high-affinity ligand of a particular protein for biotechnological applications. A so-called 'high-throughput' (HTP) SELEX experiment generates large numbers (>1000) of ligands using mass sequencing technology. We recently developed a HTPSELEX protocol, incorporating the concatemerization step of SAGE in order to increase the sequencing throughput (3). This technology development was motivated by computer simulations showing that several thousands of individual sequences are required to derive a reasonably accurate description of the sequence specificity of a typical transcription factor. Published SELEX sequence collections obtained with conventional methods rarely exceed 100 sequences.

HTPSELEX serves as a public repository for HTPSELEX data. There is a need of such a resource because the data volumes originating from HTPSELEX experiments are too large to be presented in scientific articles. In contrast, smaller SELEX sequence collections obtained with conventional methods have traditionally been disseminated through the journal literature. SELEX_DB (4) and TRANSFAC (5) are databases which offer these data in machine-readable form. Other related databases, such as JASPAR (6), only distribute the SELEX-based computational models (weight matrices) of transcription factor binding sites, but not the SELEX sequences from which these models were derived.

SCOPE AND LEADING CONCEPTS

The main purpose of the HTPSELEX database is to make the primary data from an experiment available in a form suitable for re-analysis in the future. We consider this important because the methods for characterizing the binding specificity of transcription factors are under development and continuously improving. Along with the raw data, we also make derived information available, including transcription factor binding site descriptions represented as hidden Markov

*To whom correspondence should be addressed. Tel: +41 21 692 5892/58; Fax: +41 21 652 5945; Email: Philipp.Bucher@isrec.ch models (HMMs) (7). Nevertheless, our resource is primarily intended to serve computational biologists interested in analyzing SELEX sequence libraries, either for methodological developments or for deriving better binding site models for given transcription factors.

Effective analysis of HTPSELEX data requires not only access to raw data but also precise knowledge of the experimental protocols used to generate them. A given SELEX method may introduce a specific and predictable bias in the binding site collection, which could be compensated for by a customized computational model building procedure, even though this is currently not performed in practice. The leading principle in the design of a standardized experiment description for HTPSELEX was to provide all technical details that are relevant for the downstream analysis of the data.

OVERVIEW OF A HTPSELEX EXPERIMENT

A complete SELEX experiment starts with a purified nucleic acid binding protein and terminates with a computational model of its binding specificity. Our HTPSELEX protocol, which is schematized in Figure 1, was specifically designed for DNA binding transcription factors. The transcription factor, typically, is a complex composed of several polypeptide chains produced by a recombinant organism. Note that the name of the factor is not necessarily identical with the name of one of its components deposited in the protein sequence database. Sometimes, the polypeptides used in the experiment contain only a part of the native protein. On the DNA side, the SELEX protocol uses a library of synthetic oligonucleotides consisting of a random internal part and constant flanking regions as starting material. The latter serve as PCR primers and provide restriction sites for concatemerization and insertion into a cloning vector.

Once made double-stranded, the random DNA library is mixed with protein and protein–DNA complexes are subsequently isolated by some biochemical method, for instance by preparative electrophoretic mobility shift assay. After purification, the protein–DNA complexes are dissociated and the DNA fraction is amplified by PCR before the next selection-amplification cycle. The sequences of the constant regions of the input library are relevant for downstream analysis in as much as they may overlap with the *in vitro* selected binding sites. The SELEX libraries obtained after each cycle are subjected to analytical sequencing and, if judged useful, to HTP sequencing. For this purpose, the random parts of the *in vitro* selected oligonucleotides in addition to a few flanking bases are cut out with a restriction enzyme and concatemerized before ligation into a cloning vector. Knowledge of the restriction enzyme used in this step is important as it could induce a bias in the SELEX library as binding sequences containing the corresponding restriction site are automatically destroyed during this processing step. The insert containing vectors are then transfected into bacteria. Individual colonies are sent to the sequencing facility. At this stage, the wet laboratory protocol ends and the computational data analysis pipeline starts.

The raw data obtained from the sequencing laboratory consist of trace files (electropherograms) associated with a colony identifier and a sequencing direction (forward or reverse). There may be several trace files for each colony. Individual reads from the same colony are processed and assembled with Phred and Phrap (8,9), resulting in a consensus clone insert sequence with base quality scores. Upon preliminary analysis of these sequences, one usually detects some colonies containing the same insert sequences. In the HTPSELEX jargon, these ‘colonies’ are said to represent the same ‘clone’. Sequencing reads from the same clone are pooled and subjected to a second round of Phred/Phrap processing (Figure 2).

Individual repeat units (called ‘tags’ in the HTPSELEX jargon) are parsed out from the clone sequences with the aid of a HMM representing the repetitive insert structure and some flanking vector sequences (references to HMM decoding programs are given in the next section). For each tag, a per-base error rate is computed using the base quality scores returned by Phred or Phrap. The complete tag collection is subsequently quality filtered using the error rate estimates and scanned for duplicate tags. Duplicate tags are usually observed after about five Selex cycles and are the consequence of the loss of diversity caused by repeated reduction and expansion of the population. The quality-filtered tag sequence collection is finally used to derive a binding site model. There are different types of computational models to represent the binding site, and for each type there are different algorithms to derive the corresponding parameters from the data. A survey of these methods is beyond the scope of this article. Note, however, that many of the smaller SELEX libraries published in Journal articles contain mainly high-affinity binding sites and thus are not expected to produce binding site models of high-predictive value, regardless of the model-building method used (3).

STRUCTURE AND FORMAT OF THE HTPSELEX DATABASE

The core of HTPSELEX is the flat file release, which is distributed from our FTP site jointly with the compressed archives of the trace files. There are three main files, each containing a collection of a particular entry type:

(i) htpselex.doc: contains experiment entries.
(ii) htpselex.dat: contains clone sequence entries.
(iii) htpselex.seq: contains tag sequences.

HTPSELEX entries have composite identifiers reflecting the hierarchical relationships between them. The components are alphanumeric strings separated by underscore characters. Experiment entries are identified by a short alphanumeric string, e.g. ‘NF1’ for the CTF/NF1 experiment. They contain information about the protein source, the structure of the partly random input library, the restriction enzymes used in the concatemerization step and the vector used for cloning. In addition, the number of traces, clone sequences and tags obtained from each SELEX cycle can be found there. The information is presented in a format similar to that of an EMBL or Swiss-Prot sequence entry.

The clone sequence entries contain either a complete insert sequence or a partial sequence from the 5′ or 3′ end. The latter occurs when the complete sequence of the insert could not be assembled from the sequencing reads. The clone sequence identifiers consist of the experiment ID, the cycle number,

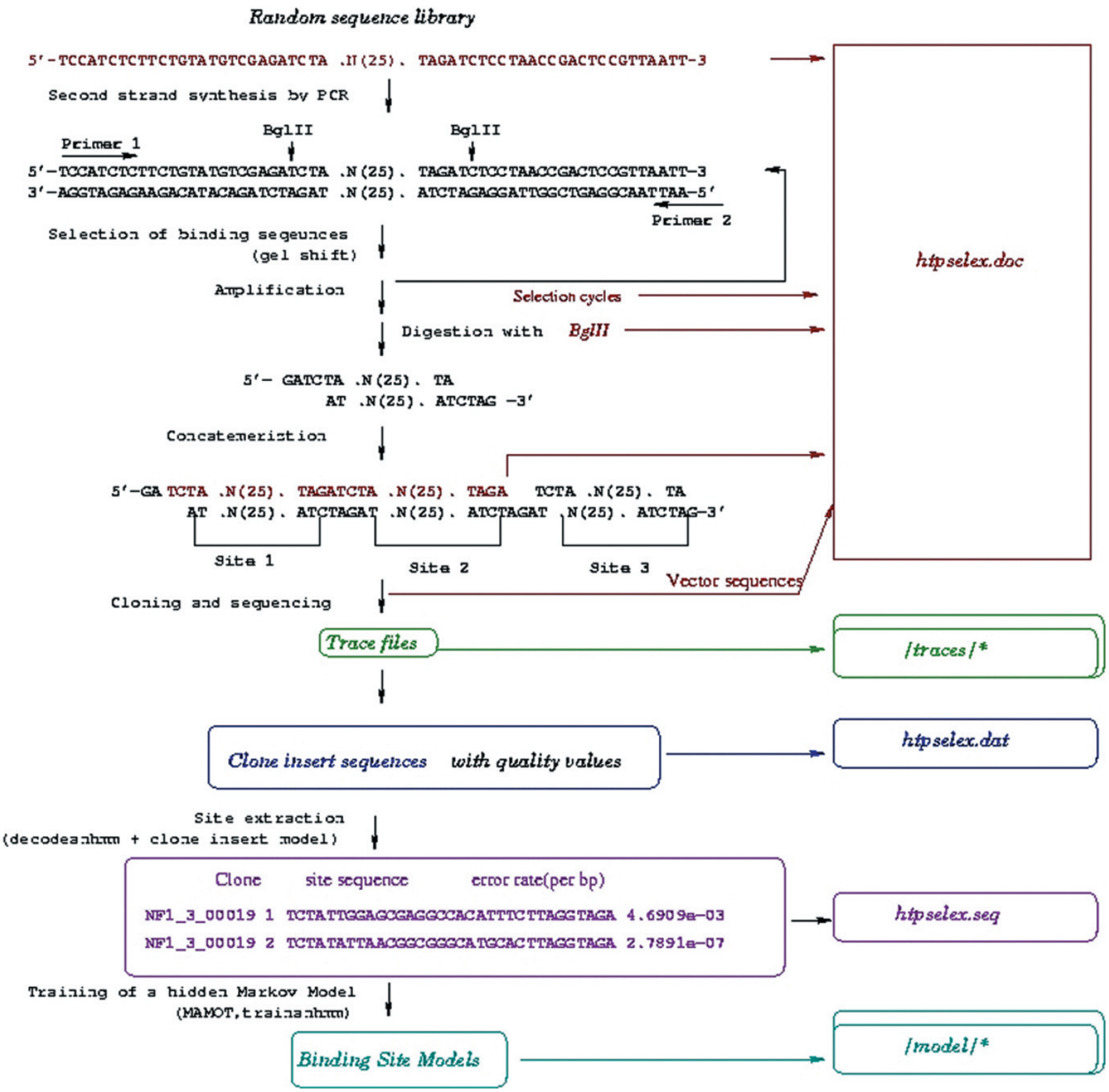
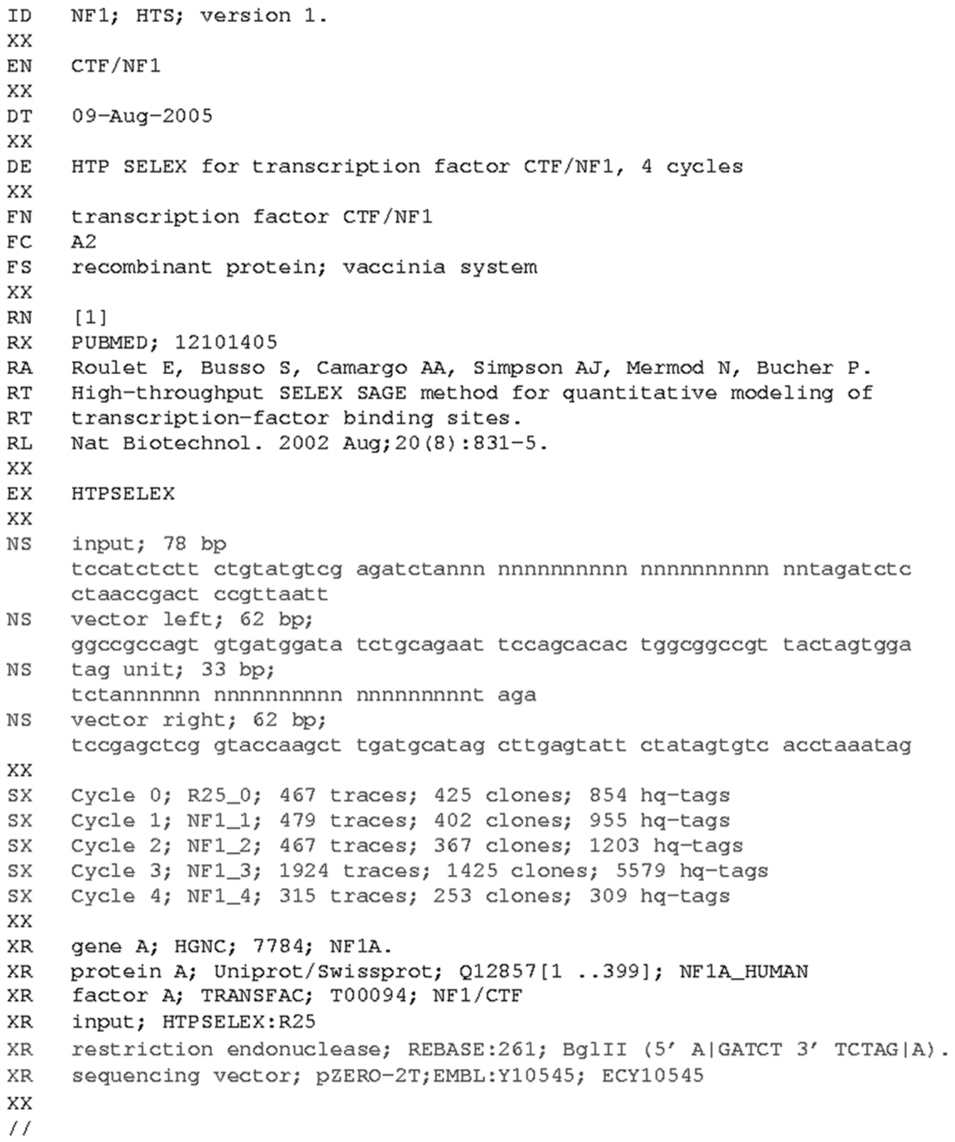

Random sequence library
5'-TCCATCTCTTCTGTATGTCGAGATCTA .N(25). TAGATCTCCTAACCGACTCCGTTAATT-3
Second strand synthesis by PCR
BglII
BglII
Primer 1
5'-TCCATCTCTTCTGTATGTCGAGATCTA .N(25). TAGATCTCCTAACCGACTCCGTTAATT-3
3'-AGGTAGAGAAGACATACAGATCTAGAT .N(25). ATCTAGAGGATTGGCTGAGGCAATTAA-5'
Primer 2
Selection of binding sequnces (gel shift)
Amplification
Selection cycles
Digestion with BglII
5'- GATCTA .N(25). TA
AT .N(25). ATCTAG -3'
Concatemeristion
5'-GA TCTA .N(25). TAGATCTA .N(25). TAGA TCTA .N(25). TA
AT .N(25). ATCTAGAT .N(25). ATCTAGAT .N(25). ATCTAG-3'
Site 1
Site 2
Site 3
Vector sequences
Cloning and sequencing
htpselex.doc
Trace files
/traces/*
Clone insert sequences with quality values
htpselex.dat
Site extraction (decodeanhmm + clone insert model)
Clone site sequence error rate(per bp)
NF1_3_00019 1 TCTATTGGAGCGAGGCCACATTTCTTAGGTAGA 4.6909e-03
NF1_3_00019 2 TCTATATTAACGGCGGGCATGCACTTAGGTAGA 2.7891e-07
htpselex.seq
Training of a hidden Markov Model (MAMOT,trainanhmm)
Binding Site Models
/model/*

Figure 1. HTPSELEX protocol. The flowchart shows the HTPSELEX data acquisition and analysis steps starting from a random pool of DNA oligonucleotides. Experimental details for each HTPSELEX experiment are recorded in the corresponding entry in htpselex.doc. The chromatograms for each experiment are made available on our FTP server. The clone sequences obtained after the Phred/Phrap processing of trace files are recorded in htpselex.dat. The 'tag' sequences corresponding to the binding sites are available in the fasta format in htpselex.seq along with quality information. Binding site models obtained after initial analysis of these experiments are also available on our FTP server.

the clone number and optionally the sequencing direction (e.g. NF1_3_00001 and NF1_3_0500_F). The feature tables are used to indicate the location of the individual tag sequences. In addition, the annotation part of these entries contains cross-references to corresponding colony names and trace files.

The tag sequences are stored in a fasta-formatted sequence file. The header line contains the tag identifier consisting of the experiment ID, cycle number, clone number and tag serial number (e.g. NF1_3_00001_1). The location in the corresponding sequence file and the estimated per-base error rate are also recorded. The tag sequence file is made non-redundant such that tags which were sequenced multiple times in the same SELEX cycle appear only once, with accessory information referring to the highest quality version.

The FTP server also provides for each SELEX cycle the trace files as a compressed archive, and a HMM representing the binding specificity of the corresponding transcription factor in two different formats suitable as input to the programs decodeanhmm (developed by Anders Krogh) and MAMOT (developed by Mauro Delorenzi, http://www.isrec.isb-sib.ch/BCF/Delorenzi/Mamot.html), respectively.



```
ID   NF1; HTS; version 1.
XX
EN   CTF/NF1
XX
DT   09-Aug-2005
XX
DE   HTP SELEX for transcription factor CTF/NF1, 4 cycles
XX
FN   transcription factor CTF/NF1
FC   A2
FS   recombinant protein; vaccinia system
XX
RN   [1]
RX   PUBMED; 12101405
RA   Roulet E, Busso S, Camargo AA, Simpson AJ, Mermod N, Bucher P.
RT   High-throughput SELEX SAGE method for quantitative modeling of
RT   transcription-factor binding sites.
RL   Nat Biotechnol. 2002 Aug;20(8):831-5.
XX
EX   HTPSELEX
XX
NS   input; 78 bp
     tccatctctt ctgtatgtcg agatctannn nnnnnnnnnn nnnnnnnnnn nntagatctc
     ctaaccgact ccgttaatt
NS   vector left; 62 bp;
     ggccgccagt gtgatggata tctgcagaat tccagcacac tggcggccgt tactagtgga
NS   tag unit; 33 bp;
     tctannnnnn nnnnnnnnnn nnnnnnnnnt aga
NS   vector right; 62 bp;
     tccgagctcg gtaccaagct tgatgcatag cttgagtatt ctatagtgtc acctaaatag
XX
SX   Cycle 0; R25_0; 467 traces; 425 clones; 854 hq-tags
SX   Cycle 1; NF1_1; 479 traces; 402 clones; 955 hq-tags
SX   Cycle 2; NF1_2; 467 traces; 367 clones; 1203 hq-tags
SX   Cycle 3; NF1_3; 1924 traces; 1425 clones; 5579 hq-tags
SX   Cycle 4; NF1_4; 315 traces; 253 clones; 309 hq-tags
XX
XR   gene A; HGNC; 7784; NF1A.
XR   protein A; Uniprot/Swissprot; Q12857[1 ..399]; NF1A_HUMAN
XR   factor A; TRANSFAC; T00094; NF1/CTF
XR   input; HTPSELEX:R25
XR   restriction endonuclease; REBASE:261; BglII (5' A|GATCT 3' TCTAG|A).
XR   sequencing vector; pZERO-2T;EMBL:Y10545; ECY10545
XX
//
```

Figure 2. Example of an experiment entry. Data items appearing in Figure 1 are shown in gray colour.

Besides HTPSELEX entries, the database also offers entries containing data from conventional SELEX experiments published in Journal articles. For convenience, these data are presented in the same format, but many fields remain empty as they are not applicable to this class. The clone insert sequence entries and trace files are missing altogether. To be acceptable for inclusion in this section, a SELEX library must contain at least 50 sequences per cycle.

The partly random input library used in our HTPSELEX experiments, was also subjected to HTP sequencing. The resulting data are contained in a special experiment entry missing all fields related to the DNA-binding protein.

So far, the HTP section of our database contains data for five different transcription factors totaling 38 254 tags. These factors are: CTF/NF1, Lef1, Lef1 in complex with β-catenin, TCF3 and TCF4. There are 26 additional entries covering conventional SELEX experiments and totaling 2278 tags. The current growth rate of the HTP section is ∼3000 tags per month and four new factors per year.

RELATIONSHIP TO OTHER DATABASES

A part of the data contained in HTPSELEX is being submitted to other databases. The trace files are currently processed by the trace archive at the NCBI (10). It has further been agreed that the tag sequences will be deposited in a special section of the EMBL database in a format similar to MGA (Mass Genome Annotation) sequences (11). The trace identifiers given by the NCBI will be cross-referenced within the clone insert sequence entries. Currently, the experiment entries contain cross-references to Swiss-Prot, EMBL (vector sequence), REBASE (12), SELEX_DB (4) and TRANSFAC (5).

ACCESS

HTPSELEX can be accessed freely via FTP (ftp://ftp.isrec.isb-sib.ch/pub/databases/htpselex) or through various web pages (http://www.isrec.isb-sib.ch/htpselex). The contents of the

FTP release has been described in detail above. The website offers as additional services:

(i) hyperlinked documentation entries for individual HTPSELEX and conventional SELEX experiments,
(ii) quality controlled download of tag sequences from individual or multiple HTPSELEX libraries with user-defined error probability thresholds,
(iii) download of tag sequences from conventional SELEX experiments and
(iv) detailed statistics for HTPSELEX experiments.

PERSPECTIVES

The HTPSELEX database is still at an early stage of its development. Several changes and extension are anticipated for the near future. A large part of the bulk data will probably soon be available from larger public data repositories at the NCBI or EBI. If this happens, the contents of HTPSELEX may be reduced to those parts not available from other sources, in the extreme case to the experiment entries only. In fact, our initiatives to submit parts of the data stored in HTPSELEX to other databases, has already stimulated a broader discussion among experts on how to store such information.

Currently, HTPSELEX contains in-house generated data and manually curated entries from Journal articles. We are, however, open to accept direct submissions from authors and are prepared to work out guidelines and automatic submission tools for this purpose in response to a demand. We are further considering the inclusion of protein-binding affinity measurements for individual oligonucleotides. Such data constitute a very useful complement to SELEX sequences for building transcription factor binding site models intended to predict the affinity of a given sequence to the protein.

ACKNOWLEDGEMENTS

The HTPSELEX database project is funded in part by grant 3100A0-104248 from the Swiss National Research Foundation. Funding to pay the Open Access publication charges for this article was provided by the Swiss Institute for Experimental Cancer Research.

Conflict of interest statement. None declared.

REFERENCES

1. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
2. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
3. Roulet,E., Busso,S., Camargo,A.A., Simpson,A.J., Mermod,N. and Bucher,P. (2002) High-throughput SELEX-SAGE method for quantitative modeling of transcription-factor binding sites. *Nat. Biotechnol*., **20**, 831–835.
4. Ponomarenko,J.V., Orlova,G.V., Frolov,A.S., Gelfand,M.S. and Ponomarenko,M.P. (2000) SELEX_DB: an activated database on selected randomized DNA/RNA sequences addressed to genomic sequence annotation. *Nucleic Acids Res*., **28**, 205–208.
5. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*., **31**, 374–378.
6. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*., **32**, D91–D94.
7. Durbin,R.M., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, Chapter 3, pp. 46–79.
8. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using *phred*. 1. Accuracy assessment. *Genome Res*., **8**, 175–185.
9. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using *phred*. 2. Error probabilities. *Genome Res*., **8**, 186–194.
10. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **33**, D39–D45.
11. Tateno,Y., Saitou,N., Okubo,K., Sugawara,H. and Gojobori,T. (2005) DDBJ in collaboration with mass-sequencing teams on annotation. *Nucleic Acids Res*., **33**, D25–D28.
12. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2005) REBASE—restriction enzymes and DNA methyltransferases. *Nucleic Acids Res*., **33**, D230–D302.