W626–W631 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl213

New Onto-Tools: Promoter-Express, nsSNPCounter and Onto-Translate

Purvesh Khatri, Valmik Desai, Adi L. Tarca[1], Sivakumar Sellamuthu, Derek E. Wildman[1], Roberto Romero[1] and Sorin Draghici*

Department of Computer Science, Wayne State University, 431 State Hall, Detroit, MI 48202, USA and [1]Perinatology Research Branch, NIH/NICHD, 4 Brush, 3990 John R, Detroit, MI 48201, USA

Received February 14, 2006; Revised and Accepted March 24, 2006

ABSTRACT

The Onto-Tools suite is composed of an annotation database and eight complementary, web-accessible data mining tools: Onto-Express, Onto-Compare, Onto-Design, Onto-Translate, Onto-Miner, Pathway-Express, Promoter-Express and nsSNPCounter. Promoter-Express is a new tool added to the Onto-Tools ensemble that facilitates the identification of transcription factor binding sites active in specific conditions. nsSNPCounter is another new tool that allows computation and analysis of synonymous and non-synonymous codon substitutions for studying evolutionary rates of protein coding genes. Onto-Translate has also been enhanced to expand its scope and accuracy by fully utilizing the capabilities of the Onto-Tools database. Currently, Onto-Translate allows arbitrary mappings between 28 types of IDs for 53 organisms. Onto-Tools are freely available at http://vortex.cs.wayne.edu/Projects.html.

INTRODUCTION

While high-throughput sequencing and microarray technologies have allowed the collection of a staggering amount of data per experiment rapidly, they have also posed the challenges of translating such data into a better understanding of the underlying biological phenomena. First released in 2001, Onto-Tools is a freely available web-accessible software suite that addresses some of these challenges (1–6). This is achieved using a probabilistic functional analysis that bridges the gap between low-level, high-throughput gene expression data and high-level functional knowledge, as well as public annotations within the framework of the Gene Ontology (GO). This analysis approach has become the *de facto* standard in the second-stage analysis of microarray experiments (7). The Onto-Tools suite includes (i) Onto-Express—used to translate lists of differentially regulated genes into a better understanding of the underlying biological phenomena; (ii) Onto-Design—used to select the best set of genes to be included on a custom microarray designed for the study of a given biological phenomenon; (iii) Onto-Compare—used to analyze the functional bias of various focused commercial microarrays and select the one that is most appropriate for a given biological hypothesis; (iv) Onto-Translate—used to translate lists of genes from one reference system to another (e.g. from GenBank accession numbers to UniGene cluster IDs to Affymetrix probe IDs, etc.); (v) Onto-Miner—providing a unified access point and an application programming interface (API) allowing queries for various information such as the gene name, official symbol, reference accession number, coded protein, etc.; (vi) Pathway-Express—which helps the users find most interesting pathway(s) involving their genes of interest; (vii) Promoter-Express—which allows the users to find condition-specific transcription factor binding sites (TFBSs) and (viii) nsSNPCounter—which allows analysis of synonymous and non-synonymous codon substitutions in protein coding genes. Previous publications have described in detail the motivation, implementation and validation of these tools (1–7). The logical work-flow between the Onto-Tools applications has been previously explained (1,4). This paper describes two new tools added to the ensemble and discusses various other additions and enhancements made to the existing tools.

PROMOTER-EXPRESS

Transcription initiation is accomplished by complex and tightly coordinated protein-DNA interactions between a number of transcription factors and the promoter region(s) of a gene. While Onto-Express and Pathway-Express in the Onto-Tools ensemble help identify the significant biological processes and pathways in the condition under study, developing a detailed mechanistic model of the regulatory mechanism(s) that control these processes requires the identification of the the genetic elements involved in these mechanisms.

*To whom correspondence should be addressed. Email: sorin@wayne.edu

Promoter-Express (PE) is a new tool in the Onto-Tools ensemble designed to help the user identify *cis*-regulatory elements on the DNA (8). The underlying hypothesis behind its approach is that similarly expressed genes involved in related biological phenomena are likely to be regulated by a common transcriptional mechanism (9–13). Given this hypothesis, PE accepts a list of genes that are known to be involved in the same or related biological processes. For example, this list could come from Onto-Express, and could contain genes involved in the same biological processes. Alternatively, the list could contain genes with similar expression profiles. For each input gene, PE queries the back-end Onto-Tools database and retrieves the nucleotide sequences in the upstream regions of all genes in the list. Currently, the organisms supported by PE are human and mouse. By default, PE retrieves a nucleotide sequence 1000 bp upstream and 200 bp downstream from the start of the mRNA of the target gene. Most TFBSs are likely to be within this region. However, PE also allows the user to expand or restrict the search boundaries. Note that in most cases, the start of an mRNA corresponds to its transcription start site (TSS). However, in case of some mRNAs, the TSS may not be annotated precisely or not annotated at all. For such situations, PE also allows the user to submit an arbitrary list of FASTA formatted sequences. Next, PE performs pairwise sequence comparisons using a sliding window approach. By default, PE uses a window size of 9 since most of the known TFBSs are 6–20 bp long. However, PE allows the user to use a different window size. Furthermore, when a match is found, the program tries to expand the matching sequence in both directions.

The result of the pairwise sequence comparisons is a number of exact matching subsequences found on both the input sequences. We define each of these exact matching subsequences as an element. It has been shown previously that two genes involved in a similar biological process and regulated by the same transcription regulation mechanism may require that these elements appear in the same order and approximately at the same distance on both genes (14). Hence, after finding the exact matching elements in both genes, PE searches for the combinations of those elements that appear in the same order with approximately the same distance among the elements on both genes. A set of elements that satisfies these criteria is defined as a module (see Figure 1). Such a module represents a 'footprint' of the transcriptional regulatory mechanisms at work in a specific biological context.

PE's output shows each input gene as a color-coded continuous line labeled with Entrez Gene ID or gene name (Figure 2). Under each line, it displays the elements found as short color-coded line segments that quickly allow the user to find out how many and what genes an element is found in. When a user moves the mouse over an element, PE shows its nucleotide sequence, start and end positions on both the genes, and the strand on which the element was found on (i.e. forward or reverse strand). Selecting a gene and one of its modules displays all elements in the selected module in color, while the rest of the elements on all genes are represented in white color (Figure 2). PE also allows the users to save the results on the user's machine in a binary file which can be opened at a later time for further analysis.

nsSNPCOUNTER

Studying the evolutionary rates of different protein coding genes usually requires the computation of the number of synonymous and non-synonymous substitutions among genes (15–17). Recent studies have focused on evolutionary changes among single nucleotide polymorphisms (SNPs) (18). Such a change is considered synonymous if it leads to a synonymous codon, i.e. a change in the nucleotide sequence of a gene does not change the amino acid sequence of the protein translated from it. Alternatively, if the change in the nucleotide sequence does change the amino acid sequence of the protein, the change is non-synonymous. Clearly, non-synonymous mutations are much more important both from an evolutionary and from a clinical perspective.

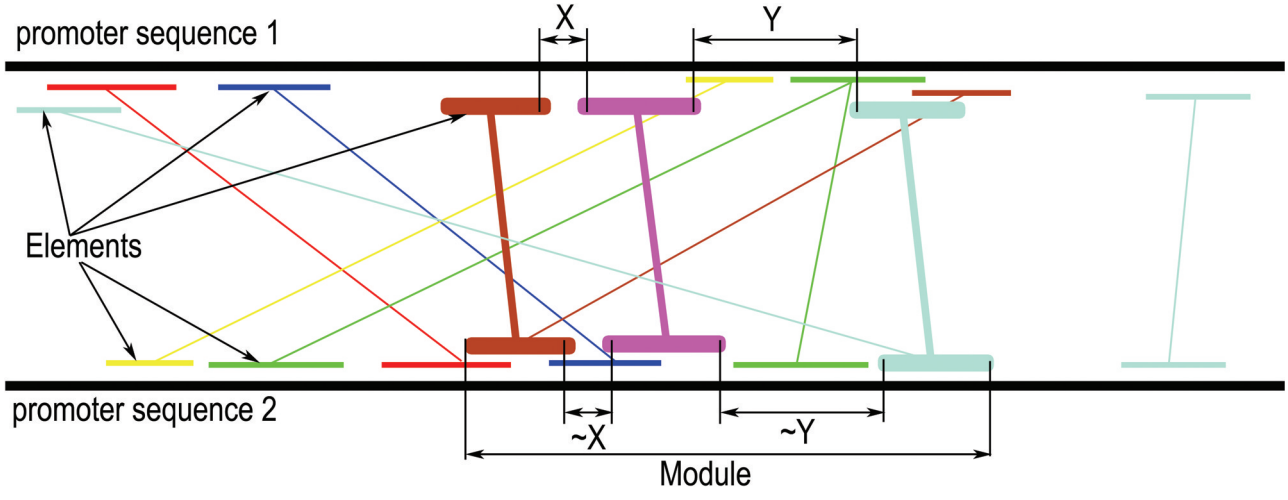

Figure 1. Example of a typical module. The black thick lines are two upstream sequences to be analyzed. The thin shorter color-coded segment between them are the elements common to both, and the thick color-coded segments are the elements that together form a module. The gap X is approximately equal to gap ~X and gap Y is approximately equal to gap ~Y.

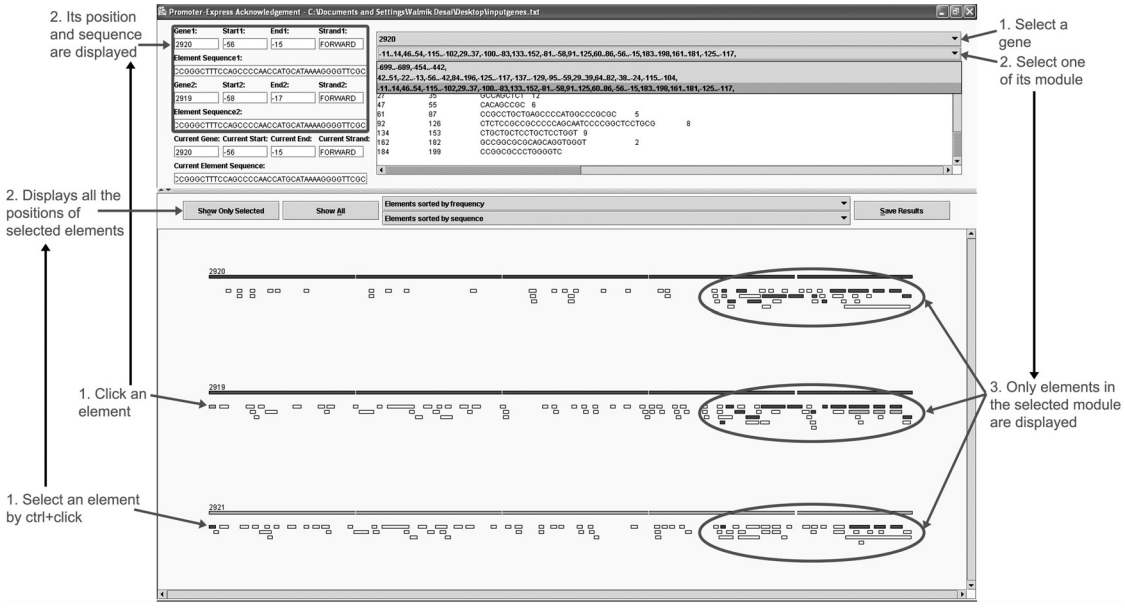

Figure 2. The output provided by Promoter-Express for a selected module. The figure also shows some of the possible data manipulations and interactions with the GUI.

The dbSNP (http://www.ncbi.nlm.nih.gov) is a SNP database provided by the NCBI that allows the retrieval of a list of known SNPs within the coding region of a given gene, identified for instance by a refseq ID. However, dbSNP does not provide the means to distinguish and automatically count the synonymous and nonsynonymous SNPs occurrences in the database (SynCounts and NonSynCounts) for a given refseq ID. The task is especially cumbersome when one needs to extract this information for thousands of genes simultaneously. The nsSNPCounter is a web-based tool that was designed to fulfil this need. Beside this main functionality it also gives the user an estimate of the number of synonymous (SynSites) and non-synonymous (NonSynSites) sites available in the sequence of each gene. This supplementary information is needed to adjust the SynCounts and NonSynCounts due to their uneven proportions (19). The PAML software collection (http://abacus.gene.ucl.ac.uk/software/paml.html) provides functionality to estimate the SynSites and NonSynSites but this is done only for one gene at a time. The new nsSNPCounter brings together all this information (SynCounts, NonSynCounts, SynSites and NonSynSites) for thousands of genes at a time.

nsSNPCounter requires a list of mRNA RefSeq IDs and the name of the organism as input. The user can also specify other optional search criteria to refine the search. These optional criteria include heterozygosity range, SNPs validation method, etc. For each RefSeq ID, nsSNPCounter queries NCBI's dbSNP database (using the *esearch* tool provided by NCBI) in order to obtain its corresponding reference SNP cluster ID (RS ID). The RS ID of the SNP is then used to obtain its corresponding gene and Entrez Gene ID using the *efetch* tool provided by NCBI. The Entrez Gene ID is further used to query dbSNP database again to retrieve all known SNPs (RS IDs) for the gene. The output of this query is further processed to retain only non-redundant SNPs, and to count the synonymous and non-synonymous substitutions relative to the reference contig.

To compute the SynSites and NonSynSites we need the sequence of the coding region of each gene of interest. This information is obtained by querying the NCBI GenBank database for every single refseq in the list to obtain a sequence GI ID. Then, the GI IDs are used to query the GenBank database again in order to retrieve the actual nucleotide sequences, and the start and end positions of the coding regions. The coding sequences are then used as an input to the PAML software which calculates the number of synonymous and non-synonymous sites. The nsSNPCounter automatically processes the PAML output and integrates the results in unique output containing the RefSeq ID, synonymous and non-synonymous SNP counts, calculated by nsSNPCounter, as well as the number of synonymous and non-synonymous sites, calculated by the PAML software (see right side of Figure 3).

ONTO-TRANSLATE

In order to correctly interpret the results of an experiment, the researchers need to build a complete picture of the biological phenomenon under study, to the extent possible, using the knowledge accumulated in various annotation databases. However, our current knowledge is spread over a number of different databases where various databases are rather specialized and no single database contains all available data. Although within each database, the data are consistent, coherent and non-redundant, most of these annotation databases are developed by independent groups. These groups use different designs and different sets of identifiers for the same biological entities. The result of these independent efforts is replication of the same information in multiple databases. Furthermore, these databases cross-reference to some of the other databases to facilitate navigation from one resource to another. In order to build a complete picture of the biological phenomenon under study, a researcher is not only responsible for mapping

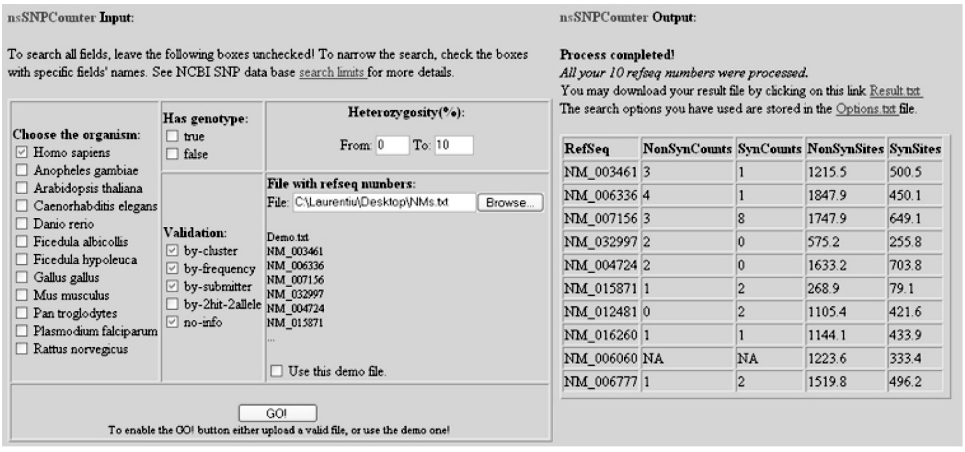

Figure 3. Input (left panel) and output (right panel) for nsSNPCounter.

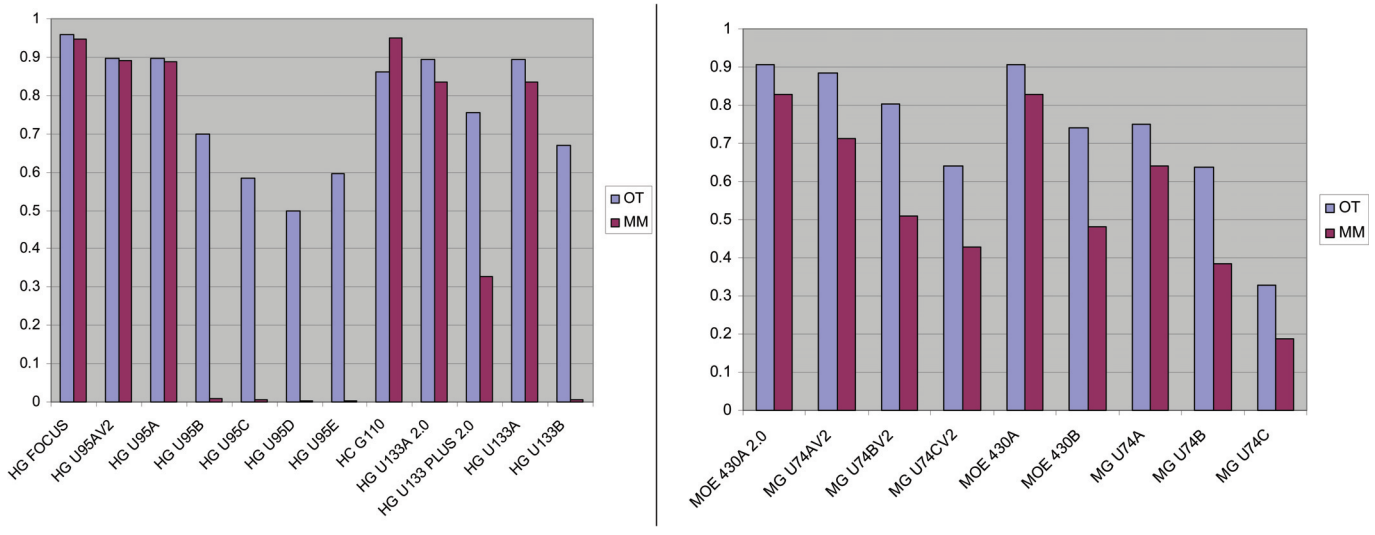

Figure 4. A comparison between the performance of Onto-Translate (OT) and MatchMiner (MM). The figures show the percentage of successful translations from probe IDs to gene symbols, for a number of sets of genes corresponding to popular Affymetrix human (left) and mouse (right) arrays.

various types of IDs from one another, but also for being aware of relationships among these resources.

Onto-Translate (OT) is designed to address these namespace issues and help the user with the problem of mapping various types of IDs to each other. The ultimate goal of OT is to provide the users with a non-redundant and complete mapping from any type of identification system to any other type. In order to achieve this goal, OT uses the custom design of Onto-Tools database that integrates 20 publicly available biological databases including dbEST (20), GenBank (21), UniGene (22), KEGG (23), WormBase (http://www.wormbase.org), NetAffx, dbEST library (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html), eVOC (24), Swiss-Prot (25), TrEMBL (25), PIR (26), UniProt (27), Eukaryotic Promoter Database (EPD) (28), Human Genome Nomenclature Committee (HGNC) (29), GenPept, Online Mendelian Inheritance in Man (OMIM) (30), Protein Data Bank (31), iProClass (26), HomoloGene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=homologene), RefSeq (32) and Gene Ontology (GO) (33,34). In addition, Onto-Tools database also integrates information about commercial microarrays from nine manufacturers including Affymetrix, Agilent Technologies, Amersham's codelink microarrays, SuperArray, Takara Biosystems, Perkin-Elmer, NIA, SigmaGenosys and Clontech. Over the past year, OT has been enhanced to allow arbitrary mappings among 28 types of IDs for 53 organisms. Currently, OT can translate thousands of IDs in a single batch run. It also provides a graphical user interface to select the desired input and output which are hyperlinked with the corresponding online database resources. As an example of the capabilities of OT, Figure 4 shows a comparison between the translations performed by OT and MatchMiner, a similar tool from NCI (35). The figure shows the percentages of genes that are successfully translated from probe IDs to gene symbols for a number of popular Affymetrix arrays. Note that NetAffx performs such translations, from probe IDs to gene symbols, but the lists to be translated are limited to at most 5000 genes. Hence, none of the translations shown here can be performed on NetAffx.

SUMMARY

The Onto-Tools suite is composed of a back-end database and eight integrated, web-accessible, free data mining tools: Onto-Express, Onto-Compare, Onto-Design, Onto-Translate, Onto-Miner, Pathway-Express, Promoter-Express and nsSNPCounter. Promoter-Express is a new tool that allows identification of condition-specific TFBSs for co-expressed genes that are involved in same or related biological processes. nsSNPCounter is another new tool that allows analysis of synonymous and non-synonymous codon substitutions for studying evolutionary rates of protein coding genes. Over the past year, Onto-Translate was enhanced to improve its scope. The Onto-Tools are freely available at http://vortex.cs.wayne.edu/Projects.html.

ACKNOWLEDGEMENTS

This work has been supported by the following grants: NSF DBI-0234806, NIH 1R01HG003491, DOD DAMD 17-03-02-0035, NIH(NCRR) 1S10 RR017857-01, MLSC MEDC-538 and MEDC GR-352, NIH 1R21 CA10074001, 1R21 EB00990-01, 1R01 NS045207-01 as well as by the Intramural Research Program of the National Institute of Child Health and Human Development, NIH, DHHS. Onto-Tools currently runs on equipment provided by Sun Microsystems under the grant EDU 7824-02344-US. Funding to pay the Open Access publication charges for this article was provided by NIH grant 1R01HG003491-01A1, Novel algorithms and organisms for Onto-Tools.

Conflict of interest statement. None declared.

REFERENCES

1. Drăghici,S., Khatri,P., Bhavsar,P., Shah,A., Krawetz,S.A. and Tainsky,M.A. (2003) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res*., **31**, 3775–3781.
2. Drăghici,S., Khatri,P., Martins,R.P., Ostermeier,G.C. and Krawetz,S.A. (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.
3. Drăghici,S., Khatri,P., Shah,A. and Tainsky,M. (2003) Assessing the functional bias of commercial microarrays using the Onto-Compare database. *BioTechniques*, 55–61.
4. Khatri,P., Bhavsar,P., Bawa,G. and Drăghici,S. (2004) Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Res*., **32**, W449–W456.
5. Khatri,P., Drăghici,S., Ostermeier,G.C. and Krawetz,S.A. (2002) Profiling gene expression using Onto-Express. *Genomics*, **79**, 266–270.
6. Khatri,P., Sellamuthu,S., Malhotra,P., Amin,K., Done,A. and Drăghici,S. (2005) Recent additions and improvements to the Onto-Tools. *Nucleic Acids Res*., **33**, W762–W765.
7. Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
8. Desai,V., Khatri,P., Done,A., Friedman,A., Tainsky,M. and Draghici,S. (2005) A novel bioinformatics technique for predicting condition-specific transcription factor binding sites. In *Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, San Diego, USA, pp. 14–15.
9. Elkon,R., Linhart,C., Sharan,R., Shamir,R. and Shiloh,Y. (2003) Genome-wide *in silico* identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Research*, **13**, 773–780.
10. Fessele,S., Maier,H., Zischek,C., Nelson,P.J. and Werner,T. (2002) Regulatory context is a crucial part of gene function. *Trends Genet*., **18**, 60–63.
11. Kel,A.E., Kel-Margoulis,O.V., Farnham,P.J., Bartley,S.M., Wingender,E. and Zhang,M.Q. (2001) Computer-assisted identification of cell cycle-related genes: New targets for E2F transcription factors. *J. Mol. Biol*.,, **309**, 99–120.
12. Long,F., Liu,H., Hahn,C., Sumazin,P., Zhang,M.Q. and Zilberstein,A. (2004) Genome-wide prediction and analysis of function-specific transcription factor binding sites. *In Silico Biol*., **4**, 395–410.
13. Sudarsanam,P., Pilpel,Y. and Church,G.M. (2002) Genome-wide co-occurrence of promoter elements reveals a *cis*-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae*. *Genome Res*., **12**, 1723–1731.
14. Werner,T., Fessele,S., Maier,H. and Nelson,P.J. (2003) Computer modeling of promoter organization as a tool to study transcriptional coregulation. *FASEB J*., **17**, 1228–1237.
15. Cargill,M., Altshuler,D., Ireland,J., Sklar,P., Ardlie,K., Patil,N., Shaw,N., Lane,C.R., Lim,E.P., Kalyanaraman,N. *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet*., **22**, 231–238.
16. Li,W.-H. (1997) *Molecular Evolution*. Sinauer.
17. Yang,Z. and Nielsen,R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol*., **17**, 32–43.
18. Bustamante,C., Fledel-Alon,A., Williamson,S., Nielsen,R., Hubisz,M., Glanowski,S., Tanenbaum,D., White,T., Sninsky,J., Hernandez,R. *et al.* (2005) Natural selection on protein-coding genes in the human genome. *Nature*, **437**, 1153–1157.
19. Yang,Z. (2002) Inference of selection from multiple species alignments. *Curr. Opin. Genet. Develop*., **12**, 688–694.
20. Boguski,M.S., Lowe,T.m.J. and Tolstoshev,C.M. (1993) dbEST—database for expressed sequence tags. *Nature Genet*., **4**, 332–333.
21. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D. (2005) Genbank. *Nucleic Acids Res*., **33**, D34–D38.
22. Schuler,G.D. (1997) Pieces of puzzle: Expressed sequence tags and the catalog of human genes. *J. Mol. Med*., **75**, 694–698.
23. Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res*., **30**, 42–46.
24. Kelso,J., Visagie,J., Theiler,G., Christoffels,A., Bardien-Kruger,S., Smedley,D., Otgaar,D., Greyling,G., Jongeneel,V., McCarthy,M. *et al.* (2003) eVOC: a controlled vocabulary for gene expression data. *Genome Res*., **13**, 1222–1230.
25. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
26. Wu,C.H., Yeh,L.-S.L., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Kourtesis,P., Ledley,R.S., Suzek,B.E. *et al.* (2003) The Protein Information Resource. *Nucleic Acids Res*., **31**, 345–347.
27. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Brigitte Boeckmann,S.F., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.
28. P′erier,R.C., Praz,V., Junier,T., Bonnard,C. and Bucher,P. (2000) The Eukaryotic Promoter Database (EPD). *Nucleic Acids Res*., **28**, 302–303.
29. Wain,H.M., Bruford,E.A., Lovering,R.C., Lush,M.J., Wright,M.W. and Povey,S. (2002) Guidelines for human gene nomenclature. *Genomics*, **79**, 464–470.
30. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*., **33**, D514–D517.

31. Berman,H.M., Westbrook,J., Feng,Z., Gilliland1,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
32. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res*., **30**, 137–140.
33. Ashburner,M. *et al*. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, **25**, 25–29.
34. Ashburner,M. *et al*. (2001) Creating the Gene Ontology Resource: Design and Implementation. *Genome Res*., **11**, 1425–1433.
35. Bussey,K.J., Kane,D., Sunshine,M., Narasimhan,S., Nishizuka,S., Reinhold,W.C., Zeeberg,B., Ajay,W. and Weinstein,J.N. (2003) Matchminer: a tool for batch navigation among gene and gene product identifiers. *Genome Biol*., **4**, R27.