PANDIT*plus*: toward better integration of evolutionary view on molecular sequences with supplementary bioinformatics resources

Slavica Dimitrieva,[1] Maria Anisimova

Department of Computer Science, Swiss Federal Institute of Technology (ETH Zurich) and Swiss Institute of Bioinformatics (SIB), Switzerland

[1]*current address:* Swiss Institute for Experimental Cancer Research (ISREC) and Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland

Abstract

Recent comparative genomic and other large-scale bioinformatics studies increasingly have been using gene annotations, functional classifications, and complementary data from the emerging "-omics" disciplines. Indeed, such analyses have better chances to uncover hidden patterns in complex multidimensional and heterogeneous biological systems data. On the other hand, inferences from such studies are extremely sensitive to data samples and quality, and are more difficult to compare or replicate owing to differences in supplementary data sources at times not publicly available. As a contribution toward the unification and integration of good quality data from heterogeneous bioinformatics resources, we present here an integrated data bank PANDIT*plus*. It is built as an extension of PANDIT, the database of PFAM alignments and phylogenetic trees for known protein domains and families spanning lineages from the three domains of life. PANDIT*plus* is a relational database containing information on functional categories, metabolic pathways, protein–protein interactions, disease associations, gene expression, three-dimensional structure, as well as estimates from evolutionary analyses of selective pressures. User-friendly interface enables customized queries and fast data access. We recommend PANDIT*plus* as a common bioinformatics platform for testing evolutionary hypotheses, which go beyond the mere inferences from molecular data by incorporating supplementary gene information. Equally, PANDIT*plus* provides an excellent resource for the development, testing, and comparison of statistical models of substitution and probabilistic dependencies between a molecular sequence and its various attributes. The database may be accessed via http://www.pandit-plus.org.

Introduction

With advances in experimental techniques, recent years observed a rapid growth not only in molecular sequence data but also in complementary gene and protein information such as gene expression, numbers of protein-protein interactions, three-dimensional structure, etc. Gene annotation equally is gaining accuracy and speed. Large-scale availability of such multi-facet data led to a common trend to incorporate the supplementary gene information with more conventional analyses of molecular sequences in order to reach more insightful conclusions. However, the results of many such studies are not easy to compare owing to their use of different data sources, produced by different laboratories, not always available to the public. Data quality and biases in gene or species samples may influence the inference. Whenever possible, it is desirable to test biological hypotheses using the same well-maintained and well-structured integrated database solution.

Here we present a relational database PANDIT*plus* that makes a step towards the integration of data from a variety of reliable and curated bioinformatics sources. Along with DNA and amino acid sequence data for homologs, PANDIT*plus* provides access to precomputed estimates from evolutionary codon models, data on protein interactions, functional and chemical pathway annotation, gene expression, and association with disease. The underlying database PANDIT[1,2] contains homologous amino acid and protein-coding sequence alignments from Pfam,[3,4] a comprehensive and accurate collection of protein domains and families.

The idea behind PANDIT database was to encourage the "evolution-centric" analyses of protein domains and families, based on reliable sets of HMM-based alignments and associated phylogenetics trees. Both Pfam and PANDIT have been updated since their first publications and became popular for large-scale studies of protein-coding genes or as testing platforms.[5-11] Among some classic examples of using Pfam/PANDIT for evolutionary model development are studies presenting novel DNA, amino acid and codon substitution models, such as WAG,[12] LG,[13] ECM,[10] SDT,[14] and THMM.[15] Recently, accuracy of the multiple protein-coding alignment method (implemented in MAGNOLIA) was tested on PANDIT data.[16] Searching for data biases and universal trends also requires large well-maintained collections of data. Multiple alignments from PANDIT and functional classification from Gene Ontology (GO) have been used to study trends relating to positive selection, and to verify and extend the complexity hypothesis.[17] Bofkin and Goldman[18] used PANDIT to demonstrate that substitution patterns at three codon positions often are strikingly different, thus necessitating suitable statistical treatment during the phylogenetic analyses.

Several authors developing statistical methodology and software for bioinformatics recently have stressed the importance of maintaining the resources like PANDIT.[11,19-21] For example, PANDIT is included as a test database by the xREI tool for phylo-grammar visualization and development,[11] which is currently a promising area in evolutionary methodology. Indeed, issues of model development, validation, and comparison may be better assessed based on a standardized data collection such as that jointly provided by PANDIT and PANDIT*plus*. The inclusion of supplementary gene information allows for better classification, filtering, and pattern discovery, contributing to the development of better statistical models. This potentially leads to greater predictive power and a better understanding of underlying evolutionary processes.

Besides automatic large-scale scans, together PANDIT and PANDIT*plus* provide a unique integrated resource for empirical case studies of sets of selected proteins. While PANDIT already has been used to study selected genes,[22-24] a wealth of supplementary gene information is readily available now via PANDIT*plus*, saving valuable time required for data mining (and, as a bonus, contributing to the reduction of web traffic). Anyone

Correspondence: Maria Anisimova, CAB H 82.1, Institute of Computational Science, ETH Zurich, Universitaetsstrasse 6, 8092 Zurich, Switzerland.
E-mail: maria.anisimova@inf.ethz.ch

Key words: comparative genomics, evolution, phylogenetics, gene family, protein domain, functional annotation.

Acknowledgements: SD was supported by the Swiss State Secretariat for Education and Research. MA was supported by the Swiss Federal Institute of Technology (ETH Zurich). We are grateful to Steven Armstrong at ETH Zurich for help with setting up the database.

Received for publication: 11 July 2009.
Revision received: 26 August 2009.
Accepted for publication: 26 August 2009.

Trends in Evolutionary Biology 2010; 2:e1
doi:10.4081/eb.2010.e1

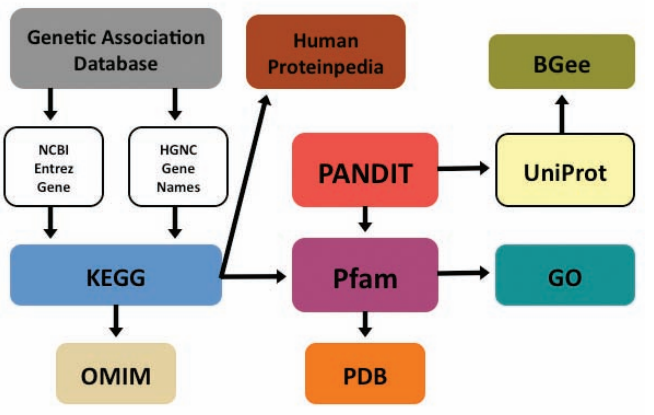

Figure 1. The mapping schema describing the integration of biological resources in PANDIT*plus*. Arrows represent the direction of the mapping between different sources. For example, an arrow leading from KEGG to Human Proteinpedia means that the mapping uses gene ID from KEGG to find associated gene information in Human Proteinpedia. For more detail on the design see the ER diagram in the Online Supplementary figure.

looking for real data examples of protein-coding genes will benefit equally from using PANDIT*plus*. Using PANDIT*plus*, subsets of genes or proteins may be selected according to defined criteria and then examined separately for a particular trend.

Structure of PANDIT*plus*

PANDIT*plus* is a MySQL database built upon the PANDIT version 17.0,[2] Protein and Associated Nucleotide Domains with Inferred Trees derived from the seed alignments of Pfam-A. Each PANDIT entry includes amino acid and codon based alignments (with reliability estimates for each column) and estimates of phylogenetic trees inferred from these alignments. The overview of biological data resources collected for each PANDIT alignment is presented in Figure 1.

Pfam at the core

Annotations for each PANDIT entry were taken from the Pfam database version 24.0.[4] Pfam is a large collection of protein families, each represented by multiple sequence alignments and profile hidden Markov models. We used Pfam-A families that each consist of a curated seed alignment containing a small set of representative members of the family, profile hidden Markov models (profile HMMs) built from the seed alignment, and an automatically generated full alignment, which contains all detectable protein sequences from the family as defined by profile HMM searches of primary sequence databases. The sequences in the Pfam-A entries are taken from the most recent releases of UniProtKB[25] and NCBI GenPept.[26] PANDIT*plus* includes general information from the Pfam database about protein families and domains with corresponding literature references, and groupings of related families and domains into clans. Information on family interactions and structural complexes that Pfam members can form as well as links to the corresponding PDB structures are taken from the Pfam database. More detailed presentation of the 3D structural information will be made in future updates.

Precomputed estimates from Markov codon substitution models

Positive or negative selection studies based on codon substitution models are powerful means of studying the evolution of genes and gene families (for review see Anisimova and Kosiol 2009).[27] Thus, we believe that maximum likelihood estimates from several standard codon models may provide illuminating details about the mode of evolution of a protein family or domain. Positive selection pressure is measured by the ratio ω of nonsynonymous to synonymous substitution rates d_N/d_S. PANDIT*plus* includes the estimates and log-likelihood scores from models M0, M1, M2, M7, and M8,[28] implemented in PAML v.4.1[29] as well as from codon models with constant and variable synonymous rates,[30] implemented in HYPHY.[31] Family entries in PANDIT*plus* are classified as positively selected or conserved based on maximum likelihood (ML) estimations and using likelihood ratio tests (LRTs) comparing couples of nested models, one of which (the null hypothesis) does not allow for positive selection (so $\omega \leq 1$), whereas another one does (the alternative hypothesis). Positive selection is detected if a model allowing sites or lineages under positive selection fits data significantly better than the model restricting selective pressure to $\omega \leq 1$ at all sites and lineages. Positions identified to be under positive selection with different models are stored in the database also. All optimizations were done assuming PANDIT trees and branch lengths were fixed to their ML estimates under the constant selective pressure model M0. Such practice commonly is used to reduce computational times, since M0 typically provides robust estimates of branch lengths. To minimize the effect of inaccuracies in automatic alignments, the analyses of selective pressure were performed only for sites where alignment was deemed reliable based on HMM analyses.[2]

The availability of precomputed results aims to increase the transparency and reproducibility of statistical analyses, and the continuity of biological studies. Note that some concerns were expressed in the literature about the suitability of codon models for evolutionary estimation on alignments of distant sequences,[32] as for deep divergences synonymous substitutions may become saturated. We therefore recommend users to discount the entries where divergence is too large; for example, average branch length >2 expected amino acid substitutions per branch. However, such high divergences are encountered rarely in PANDITplus: only eight entries fall beyond the above-mentioned threshold. Equally, we recommend a cautious use of ω estimates when either dN or d_S are close to zero (see also PAML manual). Note that Anisimova *et al*.[33] found that likelihood ratio test for positive selection remained accurate for both saturated and very similar data, although the power of the test decreased. Moreover, Seo and Kishino[34] investigated the effect of applying a codon model to divergent data sets. Surprisingly, they found that the codon model never performed worse than the amino acid model, despite large saturation of synonymous substitutions in their data.

Gene annotation

PANDIT*plus* includes gene annotation from GO and from Kyoto Encyclopedia of Genes and Genomes (KEGG). GO is a database of standardized biological terms for gene products.[35] GO contains more than 26,000 terms, divided in three branches: Molecular Function, Biological Process, and Cellular Component. Each branch can be represented as a directed acyclic graph relating terms of different degrees of specificity, with directed links from less specific to more specific terms. Each node in the graph can have several parents (broader related terms) and children (more specific related terms). GO annotations for each PANDIT member were taken

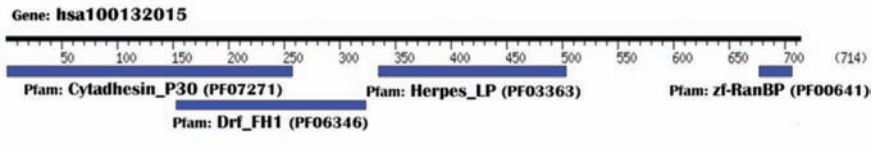

Figure 2. One-to-many mapping between a KEGG gene hsa: 100101467 and corresponding Pfam and PANDIT entries.

from the Pfam database.

KEGG is a knowledge base for systematic analysis of gene functions, linking genomic information with higher order functional information.[36,37] PANDIT*plus* integrates information from the KEGG PATHWAY database, which contains manually drawn pathway maps representing the knowledge on the molecular interaction and reaction networks. KEGG pathways are structured as a directed acyclic graph hierarchy of three flat levels. The top level consists of the following five categories: Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes, and Human Diseases. The next levels divide the five functional categories into finer subcategories. The KEGG database provides direct mapping from KEGG gene IDs to Pfam entries as well as to many other biological data sources. One KEGG gene could be mapped to many Pfam entries. An example of such mapping is illustrated in Figure 2. Each PANDIT*plus* family record contains pathway information of all the genes that are associated with the family.

Expression data

Gene expression data provide further insight into the dynamics of protein function across different species or over a series of tissues and conditions (for examples, see Akashi 2001; Khaitovich, et al. 2006; Gilad, et al. 2006; Yang, et al. 2009).[38-41] Information on expression patterns of all the genes that code for one protein or protein domain could shed light on the function of the domain in specific tissues. For example, one can evaluate the functional importance of the domain of interest in a specific tissue by observing the number of coding genes with this domain, which are expressed in that tissue. Furthermore, when species-specific expression data are available, they can be useful for studies of evolutionary patterns in genes and gene families. Together with the evolutionary model estimation, contrasting gene expression over several species or among tissues also may give insights into functional variation and evolutionary forces acting upon a gene.

PANDIT*plus* contains expression data from two recent resources, Bgee database and Human Proteinpedia, together with an overview of the number of genes expressed in various tissues. Gene expression data for five species (human, zebrafish, drosophila, mouse, and xenopus) were extracted from Bgee release 6.0.[42] Bgee focuses on the developmental aspect of gene expression and contains EST data from UniGene, Affymetrix data from ArrayExpress, and *in situ* hybridization data from ZFIN and MGI. Each datum entry is annotated manually by Bgee curators and mapped onto anatomical and developmental ontologies. The mapping to the Bgee database was done via UniProt, such that individual sequences from the seed alignment from PANDIT were mapped to UniProt IDs, and the UniProt IDs were mapped to the IDs used in Bgee, with a full path PANDIT–>UniProt–> Bgee (Figure 1). Each PANDIT*plus* family record presents tissue expression data from Bgee of the genes that code for sequences from the PANDIT seed alignment, which also was used to obtain the ML estimates from codon models (see above). Therefore, Bgee expression data are displayed only for the species that contribute a sequence in the corresponding PANDIT alignment. These expression data may be used in combination with ML estimates from phylogenetics analyses. Because one Pfam domain can be present in several genes, each with their own expression patterns, PANDIT*plus* counts the number of these genes expressed in each tissue, and sorts the tissues by the number of associated genes expressed in it. Note that interpretation of inferences based on links of gene-wise information to protein domains may not be straightforward. However, we believe that certain observations from such links may be very useful (e.g. prevalence of a domain in genes expressed in a particular tissue).

Gene expression data for healthy and disease human tissues were extracted from Human Proteinpedia, a community portal for sharing and integration of human protein data.[43,44] Human Proteinpedia allows research laboratories to contribute and maintain protein annotations. All human data in Human Proteinpedia are derived experimentally and come from different laboratories. The mapping to Human Proteinpedia gene IDs was done for the sequences from the full Pfam alignment via the KEGG database with the whole mapping path being PANDIT→PFAM→KEGG→ HumanProteinpedia (Figure 1). Each PANDIT*plus* family record presents tissue expression data from Human Proteinpedia of all the genes that are associated with the full alignment of the Pfam family. For each tissue the number of genes expressed in that tissue is displayed also. The expression data in PANDIT*plus* derived from Human Proteinpedia is only for human tissues. Since the mapping of the expression data from this resource is based on the sequences from the full alignment, such data are shown when the full alignment contains at least one human sequence, even if the seed alignment does not contain any. Owing to the fact that our Markov codon model estimates are based on the PANDIT seed alignment, we recommend that Human Proteinpedia expression data are used only when a human sequence is contained in the PANDIT seed alignment. The families that do not contain any human sequence in the PANDIT seed alignment are marked in PANDIT*plus*.

Associations with disease

PANDIT*plus* also incorporates information on associations of human genes with genetic diseases and disorders extracted from the Genetic Association Database.[45] The data in this database come from published scientific papers and this database serves as an archive of human genetic association studies of complex diseases and disorders. Note that information on associations with human diseases is available also via Online Mendelian Inheritance in Man (OMIM).[46] Links to OMIM records were taken from the KEGG database. Interpreting the disease information should be taken with caution, since one Pfam domain can be present in several genes, each linked to different diseases and disorders. Owing to this fact PANDIT*plus* counts the number of these genes per disease, and sorts the disease records by the number of domain coding genes linked with them. Furthermore, PANDIT*plus* includes literature references providing information for each gene linked to a certain disease record.

User-interface and website

PANDIT*plus* is developed with MySQL. Precomputed estimates from Markov substitution models together with other relevant data from eight different databases are organized into a compact relational database structure (*See Online Supplementary material for ER diagram*). PANDIT*plus* currently includes a total of 7738 families, corresponding to the records in the current version of PANDIT.

Table 1 shows statistics about the number of records in the current version of PANDIT*plus*.

The web interface of PANDIT*plus* was developed with PHP and JavaScript. It provides free access to the database to all academic and commercial organizations. The website of PANDIT*plus* proposes several ways to retrieve data easily. Data may be queried for families, clans, genes, gene ontologies, pathways, tissue of expression, or disease association, based on their names, identifiers, or descriptions. Quick searches for Pfam entries or clans may be done based on their accession identifiers and names, according to the nomenclature in PANDIT and Pfam. Quick search for genes may be done based on KEGG or NCBI identifiers. Protein family and clan records may be browsed based on their name. Furthermore, the web interface offers a possibility for browsing only the positively selected or conserved families under the corresponding Markov codon model, or browsing the twenty largest families in terms of number of sequences. PANDIT*plus* is accessible freely from http://www.panditplus.org/.

Table 1. PANDIT*plus* database statistics.

PANDIT alignments	7738
Domains	1800
Protein families	5637
Motifs	56
Repeats	148
Aligned sequences	181448
Families with estimates from Markov codon models	
PAML (M0, M1, M2, M7, M8)	7712
HYPHY (Dual and Non-synonymous)	7348
Interactions	5959
PDB structures	156803
Gene Ontology (GO)	
GO associations	10277
Families with at least one GO association	4286
KEGG	
KEGG pathways associations	124748
Families with at least one KEGG pathway association	4621
Number of associated human genes from KEGG	18041
KEGG pathways	310
Human Proteinpedia expression records (family – tissue)	
healthy tissues	101626
disease tissues	22611
Bgee expression records (family – tissue)	
Homo sapiens	477544
Mus musculus	298494
Drosophila melanogaster	3007
Danio rerio	2178
Xenopus tropicalis	1281
Disease information (family – disease)	
Genetic Association Database	11490
Links to OMIM	31769

The potential of PANDIT*plus* and the outlook

The database PANDIT*plus* is a powerful resource for evolutionary studies of protein domains and gene families. Clearly, the PANDIT database, once described as "apt" and "remarkably creative",[47] remains a valuable resource for evolutionary studies. PANDIT*plus* extends its potential by offering further gene and protein product information for integrated bioinformatics and molecular evolution studies. Indeed, such data become essential for the deeper exploration of the relationship between the molecular sequence and its attributes, such as sequence–function–structure, genotype–phenotype, and evolutionary rate vs. expression studies.

PANDIT*plus* has proved a useful resource for our studies of evolutionary patterns in genes and gene families, such as the investigating prevalence of positive selection, the synonymous rate variation (*Dimitrieva and Anisimova, in preparation*), and evolutionary patterns in disordered protein regions (*Szalkowski and Anisimova, in preparation*). Precomputed results from these studies will be available also from PANDIT*plus* on their publication. Expert biologists studying single genes and proteins or subclasses of proteins may find PANDIT*plus* very helpful. Extracting subsets of protein families according to defined criteria is easy with PANDIT*plus* queries. Precomputed estimates of rates from Markov substitution models may help to formulate some important selection criteria, and will be equally useful for cross comparison of results from different methods.

While our database is a useful bioinformatics resource for discovery of universal patterns and trends in protein domains and gene families, it is also an excellent test-base for the development of statistical models and methods. Soon evolutionary models will be progressing toward accommodating knowledge from the emerging "-omics" disciplines (e.g. transcriptomics, metabolomics, proteomics). Probabilistic machine learning approaches may be used to test various ideas about complex networks of interacting proteins, or to discover complex patterns. PANDIT*plus* provides an excellent ground for validating such new methods. We see further development of PANDIT*plus* to be synchronized with updates of PANDIT. We are planning to continue extending PANDIT*plus* with new estimates from statistical inferences (e.g. ML trees, ancestral state reconstruction, etc.) and new supplementary data from various sources deemed robust. In the future we aim to dedicate time and effort to the design of the automatic upgrading module, so that both PANDIT and PANDIT*plus* keep up with updates in Pfam and other supplementary databases. We actively encourage user feedback and contributions so that we can expand the range and the quality of the information currently presented in PANDIT*plus*.

Other resources that attempt to unify data from difference sources

Many recognized the need for the integration of biological resources. In fact, already some of the key database players have made steps to widen the spectrum of information covered by the repository (among these are NCBI, Pfam, UniProt, Ensembl, KEGG). However, these resources still are too specialized to contain sufficient information necessary to unify diverse molecular information with evolutionary modeling at the sequence level. Other smaller databases also attempted to integrate biological data, yet these tend to relate to particular questions or specific (smaller) samples of molecular sequences. For example, the MAGNUM database was designed to aid studies focusing on the evolution of protein structure:[48] it includes alignments mapped to crystal structure, estimates of phylogenetic trees and ancestral sequences at internal tree nodes. This information is

collected only for protein families with at least one crystal structure (around 1800). Another example is the SOURCE database that contains information on human, mouse, and rat genes, including gene ontology, functional annotations, and gene expression data.[49]

We anticipate that the number of database solutions offering the integration of resources and tools will grow rapidly in coming years. This trend is a healthy and important process for current bioinformatics and "–omics" fields, and will eventually lead to the creation of the best integrated solutions facilitating efficient data mining, data classification, and normalization.

Availability

The database is hosted on the Computational Biochemistry Research Group server and is accessible for browsing and downloads via http://www.panditplus.org/.

References

1. Whelan S, de Bakker PI, Goldman N. Pandit: a database of protein and associated nucleotide domains with inferred trees. Bioinformatics 2003;19:1556-63.
2. Whelan S, de Bakker PI, Quevillon E, et al. PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees. Nucleic Acids Res 2006;34:D327-31.
3. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 1997;28:405-20.
4. Finn RD, Tate J, Mistry J, et al. The Pfam protein families database. Nucleic Acids Res 2008;36:D281-8.
5. Brenner SE. A tour of structural genomics. Nat Rev Genet 2001;2:801-9.
6. Babu MM, Luscombe NM, Aravind L, et al. Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol 2004;14:283-91.
7. Mosavi LK, Cammett TJ, Desrosiers DC, et al. The ankyrin repeat as molecular architecture for protein recognition. Protein Sci 2004;13:1435-48.
8. Bortolussi N, Durand E, Blum M, et al. apTreeshape: statistical analysis of phylogenetic tree shape. Bioinformatics 2006; 22:363-4.
9. Gopalan V, Qiu WG, Chen MZ, et al. Nexplorer: phylogeny-based exploration of sequence family data. Bioinformatics 2006;22:120-1.
10. Kosiol C, Holmes I, Goldman N. An empirical codon model for protein sequence evolution. Mol Biol Evol 2007;24:1464-79.
11. Barquist L, Holmes I. xREI: a phylo-grammar visualization webserver. Nucleic Acids Res 2008;36:W65-9.
12. Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 2001;18:691-9.
13. Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol Biol Evol 2008;25:1307-20.
14. Whelan S, Goldman N. Estimating the frequency of events that cause multiple-nucleotide changes. Genetics 2004;167: 2027-43.
15. Whelan S. Spatial and temporal heterogeneity in nucleotide sequence evolution. Mol Biol Evol 2008;25:1683-94.
16. Fontaine A, de Monte A, Touzet H. MAGNOLIA: multiple alignment of protein-coding and structural RNA sequences. Nucleic Acids Res 2008;36:W14-8.
17. Aris-Brosou S. Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. Mol Biol Evol 2005;22:200-9.
18. Bofkin L, Goldman N. Variation in evolutionary processes at different codon positions. Mol Biol Evol 2007;24:513-21.
19. Kidd DM, Ritchie MG. Phylogeographic information systems: putting the geography into phylogeography. J Biogeogr 2006; 33:1851-65.
20. Hartmann K. Biodiversity conservation and evolutionary models. University of Canterbury, Christchurch, New Zealand, 2008.
21. Huelsenbeck JP, Joyce P, Lakner C, et al. Bayesian analysis of amino acid substitution models. Philos Trans R Soc Lond B Biol Sci 2008;363:3941-53.
22. Massingham T, Goldman N. Detecting amino acid sites under positive selection and purifying selection. Genetics 2005; 169:1753-62.
23. Collins K, Gu H, Field C. Examining protein structure and similarities by spectral analysis technique. Stat Appl Genet Mol Biol 2006;5:Article23.
24. Morrison DA. A framework for phylogenetic sequence alignment. Plant Syst Evol 2008. DOI 10.1007/s00606-008-0072-5.
25. The UniProt Consortium. The Universal Protein Resource (UniProt). Nucleic Acids Res 2008;36:D190-5.
26. Burks C, Cassidy M, Cinkosky MJ, et al. GenBank. Nucleic Acids Res 1991;19: S2221-5.
27. Anisimova M, Kosiol C. Investigating protein-coding sequence evolution with probabilistic codon substitution models. Mol Biol Evol 2009;26:255-71.
28. Yang Z, Nielsen R, Goldman N, et al. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 2000;155:431-49.
29. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 2007; 24:1586-91.
30. Kosakovsky Pond SL, Muse SV. Site-to-site variation of synonymous substitution rates. Mol Biol Evol 2005;22:2375-85.
31. Kosakovsky Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics 2005;21:676-9.
32. Maynard Smith J, Smith NH. Synonymous nucleotide divergence: what is saturation? Genetics 1996;142:1033-6.
33. Anisimova M, Bielawski JP, Yang Z. Accuracy and power of the likelihood ratio test to detect adaptive molecular evolution. Mol Biol Evol 2001;18:1585-92.
34. Seo T-K, Kishino H. Synonymous substitutions substantially improve evolutionary inference from highly diverged proteins. Syst Biol 2008;57:367-77.
35. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat Genet 2000;25:25-9.
36. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27-30.
37. Kanehisa M, Goto S, Hattori M, et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 2006;34:D354-7.
38. Akashi H. Gene expression and molecular evolution. Curr Opin Genet Dev 2001;11: 660-6.
39. Khaitovich P, Enard W, Lachmann M, et al. Evolution of primate gene expression. Nat Rev Genet 2006;7:693-702.
40. Gilad Y, Oshlack A, Smyth GK, et al. Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature 2006;440:242-5.
41. Yang D, Jiang Y, He F. An integrated view of the correlations between genomic and phenomic variables. J Genet Genomics 2009;36:645-51.
42. Bastian F, Parmentier G, Roux J, et al. Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. Data Integration in the Life Sciences, 2008.
43. Mathivanan SM, Ahmed NG, Ahn H, et al. Human Proteinpedia enables sharing of human protein data. Nat Biotechnol 2008;26:164-7.
44. Kandasamy K, Keerthikumar S, Goel R, et al. Human Proteinpedia: a unified discovery resource for proteomics research. Nucleic Acids Res 2009;37:D773-81.
45. Becker KG, Barnes KC, Bright TJ, et al. The genetic association database. Nat

Genet 2004;36:431-2.
46. McKusick VA. Mendelian Inheritance in Man and its online version, OMIM. Baltimore: Johns Hopkins University Press, 1998.
47. Galperin MY. The Molecular Biology Database Collection: 2005 update. Nucleic Acids Res 2005;33:D5-24.
48. Bradley ME, Benner SA. Phylogenomic approaches to common problems encountered in the analysis of low copy repeats: the sulfotransferase 1A gene family example. BMC Evol Biol 2005;5:22.
49. Diehn M, Sherlock G, Binkley G, et al. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Res 2003;31:219-23.