D724–D730 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj045

Phytome: a platform for plant comparative genomics

Stefanie Hartmann, Dihui Lu[1], Jason Phillips and Todd J. Vision*

Department of Biology and [1]School of Information and Library Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

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

ABSTRACT

Phytome is an online comparative genomics resource that can be applied to functional plant genomics, molecular breeding and evolutionary studies. It contains predicted protein sequences, protein family assignments, multiple sequence alignments, phylogenies and functional annotations for proteins from a large, phylogenetically diverse set of plant taxa. Phytome serves as a glue between disparate plant gene databases both by identifying the evolutionary relationships among orthologous and paralogous protein sequences from different species and by enabling cross-references between different versions of the same gene curated independently by different database groups. The web interface enables sophisticated queries on lineage-specific patterns of gene/protein family proliferation and loss. This rich dataset is serving as a platform for the unification of sequence-anchored comparative maps across taxonomic families of plants. The Phytome web interface can be accessed at the following URL: http://www.phytome.org. Batch homology searches and bulk downloads are available upon free registration.

INTRODUCTION

The comparative analysis of genome data can provide unique and valuable insights into organismal function and evolution. While a treasure trove of data is publicly available, there are substantial barriers to the exploitation of these data in a typical small-scale research project. Data often come from many different sources, each with different conventions for data management. This problem is particularly pronounced in plants owing to the highly decentralized infrastructure for plant genomics. In addition, comparative genomic analyses frequently require complex and computationally intensive software not accessible to the typical lab, and occasional users are unaware of the state of the art. To address these issues, we have developed Phytome, an online comparative genomics resource for functional plant genomics, molecular breeding and evolutionary studies. Phytome centralizes the relevant data and makes the results of its computationally intensive analysis pipeline available through a versatile and powerful web-based graphical user interface (GUI), thereby enabling individual researchers to utilize the tools of plant comparative genomics. In its current form, Phytome is well suited to studies of functional diversification of protein-coding gene families and taxonomic lineages. In addition, Phytome serves as glue between otherwise disjointed plant unigene databases and between taxon-specific model organism databases.

Phytome contains publicly available protein sequence information from a phylogenetically diverse set of plant species (Figure 1). Thirty-nine taxa are included in version 1 (released September 2004) and over one hundred taxa are, at the time of writing, planned for inclusion in version 2 (to be released Fall 2005). The majority of protein sequences are computationally predicted from expressed sequence tag (EST) data, but Phytome also includes protein predictions from genomic DNA and full-length cDNA sequences when available. The sequence data are updated on an annual basis and new features are added with each release. The most significant new functionality anticipated in forthcoming releases is the inclusion of comparative mapping data along with novel analysis and visualization tools for comparative maps.

DATABASE CONTENT AND ANALYSIS PIPELINE

In brief, the multistage analysis pipeline (Figure 2) begins with whole or partial predicted protein-coding genes (Unigenes). From these, predicted translations (Unipeptides) are obtained. Unipeptides are grouped into Unipeptide Families, and multiple sequence alignments (MSAs) and phylogenies are inferred for each family. Large families are then further broken down into phylogenetically-defined Unipeptide Subfamilies. A variety of functional annotation tools are applied to characterize and classify representative sequences from each Unipeptide Subfamily. Here, we describe each step of the pipeline in turn. We address differences between version 1 and 2 of Phytome where relevant but primarily focus on the analysis pipeline for version 2 (under development) and

*To whom correspondence should be addressed. Tel: +1 919 843 4507; Fax: +1 919 962 1625; Email: tjv@bio.unc.edu

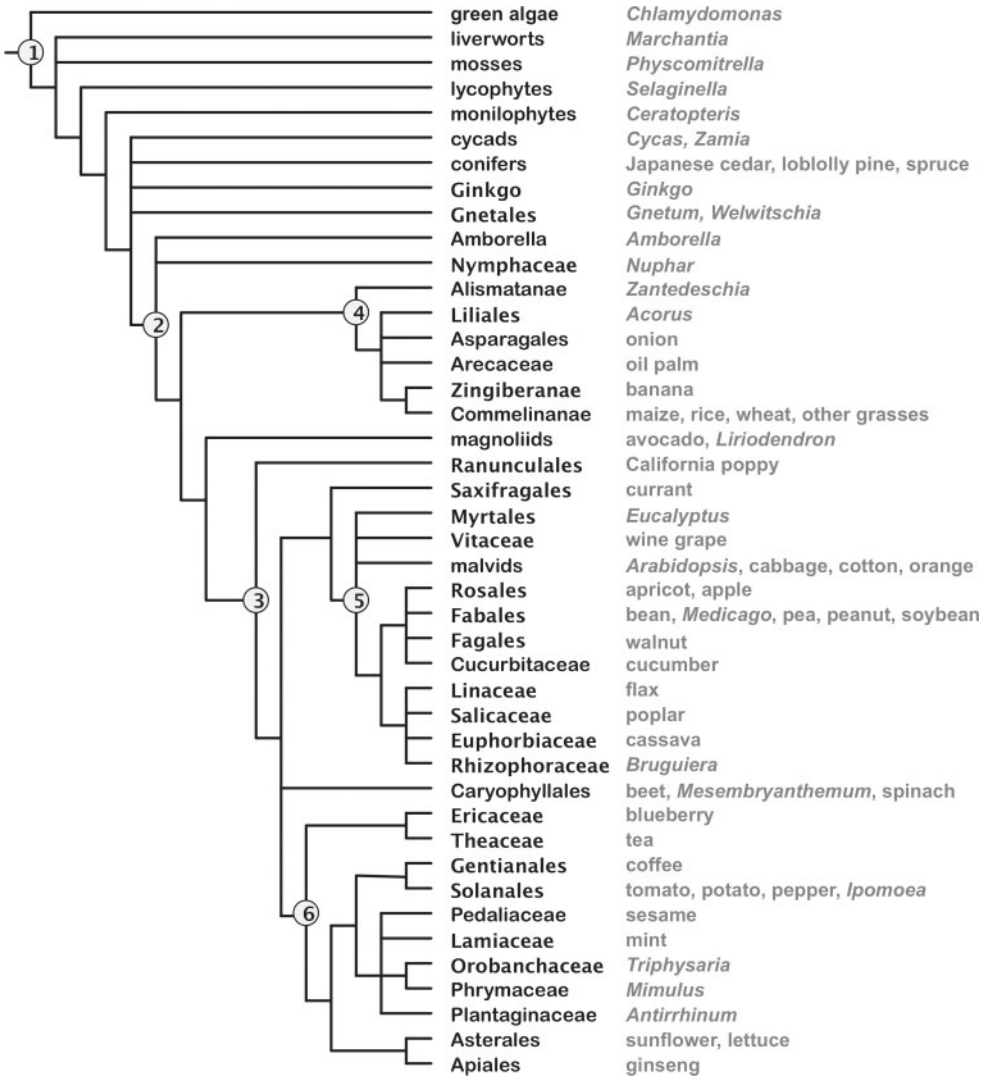

Figure 1. Phylogeny of plant model organisms. Lineages represented in Phytome are shown. Representative species are denoted by common name or genus. The tree is largely based on recommendations by the Angiosperm Phylogeny Group (APG II). Major groups are indicated by numbered circles at the nodes of their last common ancestor: 1, Viridiplantae; 2, angiosperms; 3, eudicots; 4, monocots; 5, rosids; 6, asterids.

provide statistics for version 1. More information about implementation of the analysis pipeline is available at the website, and a detailed evaluation of the results will be published elsewhere.

Unigenes

For both efficiency and accuracy in the pipeline, Phytome stores one or more non-redundant sets of predicted genes (a Unigene set), for each species. We distinguish between a primary Unigene set, composed of the Unigenes used for subsequent sequence analyses, and secondary Unigene sets, for which only assembly information is stored so that corresponding Unigenes from different sources can be cross-referenced. Pre-assembled Unigene sets for many of the species are obtained from NCBI Unigene (1), Plant Genome Network (PGN) (2), PlantGDB (3), Sputnik (4) and TIGR Gene Indices (5). We prefer to use as a primary source a Unigene set built using base call quality data [e.g. Phred scores, refs. (6,7)], though this is not possible for many of the species. When such an assembly is available from the original data provider, an effort is made to obtain that assembly rather than one built without quality scores. For species without an existing Unigene set in the public domain, one is custom-built using the software TGICL and CAP3 (8,9). When an authoritative genome annotation is available for a species (e.g. *Arabidopsis*, rice), the primary Unigene set is obtained from the corresponding genome annotation group.

Unipeptides

In most cases (*Arabidopsis* and rice excepted), Unipeptides must then be inferred from the Unigene sequences. A multi-stage homology search is done against several protein sequence databases using BLAST (10,11). First, Uniprot/Swissprot plus TrEMBL plant proteins are searched. If a nearly perfect match is found to a protein from the same species, this protein (or a consensus of all such proteins) is

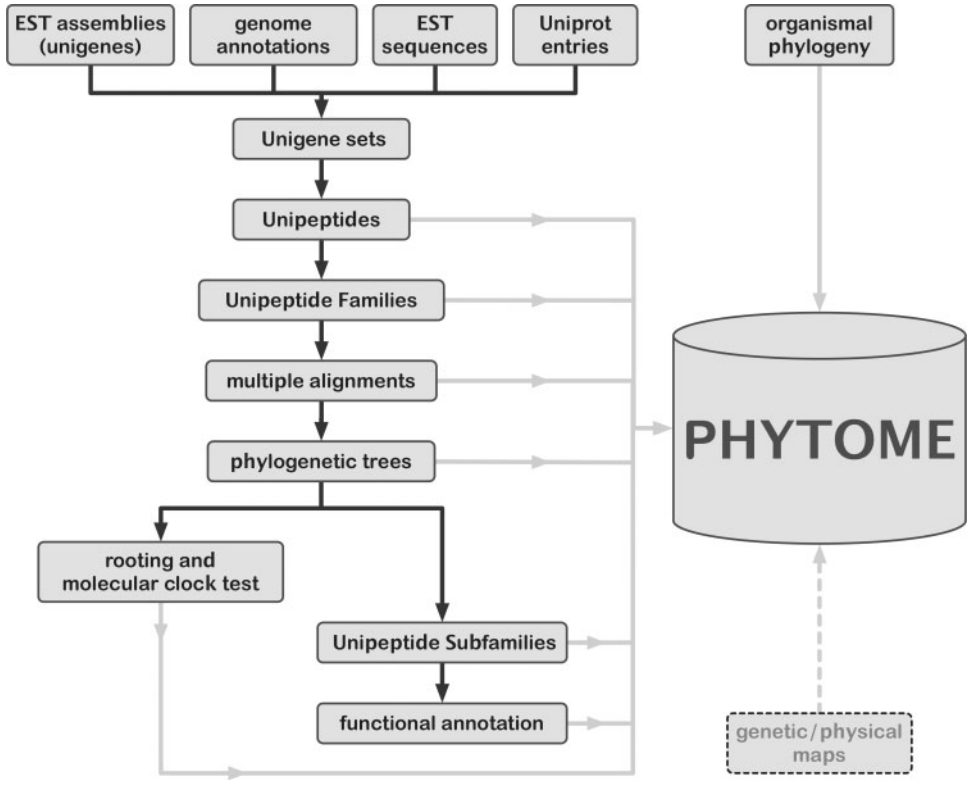

Figure 2. Phytome analysis pipeline A simplified schematic representation of the Phytome version 2 analysis pipeline. Black arrows represent the flow of data. Light gray arrows indicate output stored in the database. Integration with data from physical and genetic maps (dashed box and arrow) is planned for future releases.

used as the Unipeptide for that Unigene. Failing that, the top three homologs are input to a homology-guided translation step using ESTWise (12) using the following three datasets in descending order of priority: (i) all Uniprot/Swissport plus TrEMBL plant records, (ii) non-plant records in Uniprot/Swissprot, or (iii) non-plant records in Uniprot/TrEMBL. Some Unigenes do not produce a corresponding Unipeptide in Phytome for any number of (non-mutually exclusive) reasons: they may lack a coding sequence (by consisting entirely of the 5′ or 3′-untranslated region, or of an RNA gene), they possess a coding sequence that is too short, homologs can not be found, or the homology-based translation fails. In version 1 of Phytome, there were 735 024 Unigenes but only 640 467 Unipeptides, an attrition of ∼13%. Basic information, such as component sequences, can still be retrieved even for a Unigene lacking a corresponding Unipeptide.

Unipeptide Families

In version 1 of Phytome, an all-by-all BLASTP of Unipeptides from all species was used as input to Tribe-MCL (13), which outputs non-overlapping clusters of sequences that can be considered approximate Unipeptide Families. Tribe-MCL heuristically takes into account both the strength of the pairwise matches and the interconnectivity among the members of a cluster. The 'inflation value', which is a tunable parameter affecting the stringency of the clustering, was initially set to three. Clusters with >400 members were iteratively broken into smaller clusters using inflation values of four or five. In version 1, clustering was initiated with 640 467 Unipeptides. Following clustering, there were 26 393 Unipeptide Families of size two or greater. In addition to these, there were 94 537 Unipeptides with no BLAST match and 212 562 Unipeptides that were not included in any cluster despite having BLAST matches of $E \leqslant 10^{-15}$. Thus, ∼48% of the Unipeptides were grouped into a family containing at least one other Unipeptide.

For subsequent versions of Phytome, the process of family assignment builds upon existing families as much as possible. Phytome version 2 Unipeptides that have not changed in sequence retain their earlier family membership. New and updated Unipeptides are first searched against profile hidden Markov models (HMMs) generated using HMMer (14) for the large Unipeptide Families from version 1. Those that clearly fall into existing families need not be clustered. Those that do not match an existing profile HMM are searched against version 1 families that were too small to have HMMs, and against each other, using BLAST. They are assigned to one of the small families if they are closer in sequence to one member than current family members are to each other. If not, they are clustered into new Unipeptide Families as in version 1.

Alignments

A MSA is produced for every Unipeptide Family. Owing to the well-known difficulties of automated *de novo* MSA on large protein families, especially when there are many incomplete sequences, different software programs and parameters need to be applied depending on the context. The vast majority of *de novo* MSAs for Phytome version 1 have been produced using MAFFT, which is both extremely rapid and has been

shown to produce alignments comparable with the most successful general-purpose programs available (15,16). For each family, MAFFT was run once with two iterations and once with three; the better alignment, as determined by the average sum of pairs (SP) score with the PAM250 substitution matrix (12), was retained. Alignments of families with 20–700 members generated by MAFFT were subsequently refined using RASCAL (17), but since we encountered cases in which applying RASCAL yielded a lower quality MSA than was input, the refined alignment was used only when it had a higher SP score. For a small number of families, the MAFFT alignment was found to be inadequate, and T-COFFEE or DIALIGN (18–20) were used to generate alignments for these. To ensure positional homology of columns used for reconstructing phylogenies (see below), we use a custom program named REAP (S. Hartmann, J. Phillips, T.J. Vision, unpublished data) to 'prune' the full MSA by removing (i) columns containing many gaps and/or highly diverse amino acids and (ii) sequences that either have little overlap with other sequences or appear to be systematically misaligned. The output of REAP is referred to in Phytome as the 'reduced alignment'.

As discussed above, profile HMMs were trained for Unipeptide Families of Phytome version 1 and used to identify new family members, using full alignments. Profile HMMs were also used to guide full alignments of version 2 Unipeptide Families using HMMer, rather than recalculating each MSA from scratch. Families for which no prior profile HMM is available are aligned as in version 1. Profiles are calculated for new families and for old ones that change substantially in membership between versions.

Phylogenies

From reduced alignments of Unipeptide Families with at least four sequences, phylogenies are calculated using the Neighbor-Joining algorithm (21). Pairwise distances (using the JTT matrix) are calculated using PHYLIP Protdist and input into PHYLIP Neighbor to obtain an unrooted phylogenetic tree (J. Felsenstein, University of Washington, Seattle, WA, USA). To facilitate downstream applications such as inferring speciation and duplication events, all trees are midpoint-rooted using PHYLIP Retree; a molecular clock test is then performed to determine the reliability of this root. The reduced alignment and neighbor-joining tree for each Unipeptide Family are read into TreePuzzle (22), and a likelihood ratio clock test is performed (also using the JTT model). In version 1 of Phytome, phylogenetic trees were calculated for 11 390 Unipeptide Families. A molecular clock was rejected for 8111 (71%) of the trees.

Unipeptide Subfamilies

Subfamilies are identified to facilitate analysis within large families. They are solely defined by the phylogenetic structure of the family; no functional information is taken into account. To obtain Unipeptide Subfamilies, each midpoint-rooted tree is traversed by a breadth-first search from the leaves to the root. During this traversal, monophyletic clades containing up to 50 leaves are selected and defined as Unipeptide Subfamilies. Sequences excluded by REAP from the MSA, and therefore the phylogenetic tree, are placed into a special Unipeptide Subfamily (numbered 0).

Functional annotations

A number of automated functional predictions are used to characterize and classify the Unipeptides within Phytome. Chloroplast and mitochondrial encoded Unipeptides are identified within several of the Phytome species using BLAST searches for near perfect matches against predicted proteins from the completed organelle genomes of those same taxa. InterproScan (23,24) is used to predict domains and functional motifs for the longest Unipeptides within each Subfamily. This performs searches for conserved signatures within the following protein domain and motif databases: Interpro (25,26), PIR Superfamily (27–29), PRINTS (30), PROSITE/Prefile (31), PFAM (32), PRODOM (33,34), SMART (35), Superfamily (36) and TIGRFAMs (37). Low-complexity regions as determined by SEG (38) are also noted. GO (39) terms are assigned to Unipeptides on the basis of the InterPro2GO mapping, which is generated during the curation of InterPro entries (26). Signal peptides and transmembrane domains are predicted using SignalP (40,41) and TMHMM (42), respectively. Since protein function may often be conserved within a Unipeptide Subfamily, and since these predictions tend to be computationally expensive, only one representative of each Subfamily is analyzed by InterproScan, SignalP and TMHMM.

Implementation

The results of the analysis pipeline are stored in relational tables using a custom schema. Phytome runs on a MySQL backend, and the web GUI consists of dynamic HTML documents generated by PHP. The custom middle layer makes heavy use of Bioperl (43).

HOW TO USE THE DATABASE

Phytome's web-based GUI allows individual users intuitive access to the contents of the database, along with a variety of visualization tools. Registered users can download table dumps, which are intended to encourage connectivity with Phytome from external databases and websites. Registration is free and can be accessed from Phytome's Advanced Features page.

Unipeptides

Unipeptides can be retrieved in a number of different ways. Each has a unique Unipeptide ID, which consists of a four letter species code followed by an integer. A unipeptide can also be retrieved by querying one of its component sequences (including a Genbank EST, Uniprot ID, a unigene ID from another database, or a gene model identifier such as an *Arabidopsis* At or rice Os number). Synonyms for Unipeptides are stored as searchable gene/marker aliases. The database of synonyms is not comprehensive, but seeks to include synonyms particularly when they are used for markers on genetic or physical maps. It is also possible to retrieve all Unipeptides to which a particular Interpro or GO ID or term has been assigned, although such assignments are only available for one exemplar in each Unipeptide Subfamily. All searches may be restricted by species.

A unique feature of Phytome is the ability to cross-reference unigenes from external databases whose assemblies include one or more shared component sequences. This information is displayed in two fields on the Unipeptide page: 'Primary source' and 'Secondary sources'. The primary source is the DNA sequence that was used to obtain the Unipeptide sequence and used for all sequence analyses. The secondary sources share one or more component sequences (usually ESTs) with the primary source. This cross-reference is possible because Phytome stores the Genbank accession numbers of the component sequences for Unigenes, at least for those Unigene assemblies for which these data are available.

One can enter a query DNA or protein sequence to perform a BLAST search for homologous Unipeptides within a single species or all species combined. The customized BLAST output organizes the hits in a novel way, by Unipeptide Family. Batch BLAST is also available to registered users.

Unipeptide Families

Unipeptide Families can be directly retrieved similarly to Unipeptides, by entering a text string or using BLAST. Each family has a unique numerical ID. With a few exceptions, the smaller the family number, the more Unipeptides in that family. In addition, families can be retrieved by searching for Interpro and GO IDs and terms that have been assigned to Unipeptides within a given family. The list of Unipeptides within a family can be viewed sorted either by Unipeptide Subfamily or by species. A graphical depiction of the Interpro (and GO) assignments for all Subfamilies within the larger Unipeptide Family can be obtained. Related Unipeptide Families are also listed. The user then selects all or a subset of Unipeptides from a family to include in a multiple alignment and/or phylogeny.

A novel way to browse Unipeptide Families is to search for those that do or do not contain members from particular species or clade using a graphical Species Selector tool. With this tool, one can require a species to be either present or absent; since species with small numbers of Unipeptides will necessarily lack members in most families, requiring a species to be present can sometimes be too restrictive for a particular search.

Alignments and phylogenies

Once the user selects which Unipeptides within a family to include in the alignment, a page displaying the reduced alignment is shown and a set of customized unigene assembly, sequence, alignment and phylogeny files are made available for viewing and download. The alignment and phylogeny of the selected Unipeptides can be viewed and manipulated interactively using the JalView (44) and ATV (45) Java applets, respectively.

Examples

Two brief examples of published studies will help to illustrate how seemingly complex comparative genomics questions can be answered quite easily using Phytome. Allen (46) investigated the role of gene loss and gene acquisition in angiosperms. To find genes that had been lost in the lineage leading to *Arabidopsis*, the author ran BLAST against custom-built unigene assemblies. He compiled a list of 1002 tomato genes that lacked an *Arabidopsis* homolog. He then selected the 154 genes from that list that had a match in either soybean or *Medicago*. Phytome's Species Selector can be used to query for Unipeptide Families that show particular patterns of lineage-specific presence and absence. As expected, a quick Phytome search designed to reproduce Allen's results retrieves the genes discussed in the original study (e.g. polyphenol oxidases, ornithine decarboxylases and cyanobacterial proteins), in addition to many more, conveniently grouped by Unipeptide Family.

Phytome is also especially well suited for the study of orthology and paralogy within large gene families. In a recent study of receptor-like kinases in *Arabidopsis* and rice (47), the authors identified and retrieved kinase sequences and used these to generate an MSA. The phylogenetic tree computed from the alignment was used to identify (i) ancestral family members present in the common ancestor and (ii) subfamilies that show differential expansion between the species. With Phytome, Unipeptide Families can be retrieved using keywords based on InterPro and GO terms (e.g. 'kinase'). The user can then display a phylogeny for all the Unipeptides within a family from selected species, facilitating analysis of diversification within any of the thousands of Unipeptide Families in the database.

FUTURE DIRECTIONS

Phytome is currently being expanded to allow comparisons among maps from distantly related plant species. This will be done by first storing the correspondence between Unipeptides already in Phytome and sequence-based markers on genetic maps, physical maps and assembled genome sequences from those species for which such information is available [see also ref. (48)]. By joint analysis of the phylogenetic relationships among mapped Unipeptides and the genetic/physical locations of Unipeptide Family members in each species, segments of chromosomal homology among multiple species will be inferred and recorded in the database. This will allow next-generation tools that can, for instance, infer the gene content of a chromosomal segment from a sparsely-mapped plant genome, provided sequenced markers have been mapped to that region in sufficient density to identify syntenic segments in genomes with denser marker coverage. The inclusion of increasingly sensitive methods for detecting synteny in highly diverged chromosomal segments (49,50), comparative methods for predicting gene content and organization in sparsely sampled genomes and visualization tools for the results of these analyses will put automated predictions of gene content in candidate QTL regions at experimentalists' fingertips.

ACKNOWLEDGEMENTS

The authors thank V. Brendel, L. Mueller, S. Rudd and all the other data providers upon whom Phytome depends. This work is supported by a grant from the National Science Foundation (DBI-0227314). Funding to pay the Open Access publication charges for this article was provided by the NSF and the University of North Carolina.

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Albert,V.A., Soltis,D.E., Carlson,J.E., Farmerie,W.G., Wall,P.K., Ilut,D.C., Solow,T.M., Mueller,L.A., Landherr,L.L., Hu,Y. *et al.* (2005) Floral gene resources from basal angiosperms for comparative genomics research. *BMC Plant Biol.*, **5**, 5.
3. Dong,Q., Schlueter,S.D. and Brendel,V. (2004) PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res.*, **32**, D354–D359.
4. Rudd,S. (2005) openSputnik—a database to ESTablish comparative plant genomics using unsaturated sequence collections. *Nucleic Acids Res.*, **33**, D622–D627.
5. Lee,Y., Tsai,J., Sunkara,S., Karamycheva,S., Pertea,G., Sultana,R., Antonescu,V., Chan,A., Cheung,F. and Quackenbush,J. (2005) The TIGR gene indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res.*, **33**, D71–D74.
6. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome. Res.*, **8**, 186–194.
7. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome. Res.*, **8**, 175–185.
8. Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome. Res.*, **9**, 868–877.
9. Pertea,G., Huang,X., Liang,F., Antonescu,V., Sultana,R., Karamycheva,S., Lee,Y., White,J., Cheung,F., Parvizi,B. *et al.* (2003) TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.
10. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
11. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
12. Birney,E., Clamp,M. and Durbin,R. (2004) GeneWise and Genomewise. *Genome. Res.*, **14**, 988–995.
13. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
14. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
15. Katoh,K., Kuma,K., Toh,H. and Miyata,T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
16. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
17. Thompson,J.D., Thierry,J.C. and Poch,O. (2003) RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics*, **19**, 1155–1161.
18. Morgenstern,B., Frech,K., Dress,A. and Werner,T. (1998) DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics*, **14**, 290–294.
19. Morgenstern,B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **315**, 211–218.
20. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **1302**, 205–217.
21. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
22. Schmidt,H.A., Strimmer,K., Vingron,M. and von Haeseler,A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**, 502–504.
23. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
24. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
25. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
26. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
27. Wu,C.H., Huang,H., Nikolskaya,A., Hu,Z. and Barker,W.C. (2004) The iProClass integrated database for protein functional analysis. *Comput. Biol. Chem.*, **28**, 87–96.
28. Wu,C.H., Nikolskaya,A., Huang,H., Yeh,L.S., Natale,D.A., Vinayaka,C.R., Hu,Z.Z., Mazumder,R., Kumar,S., Kourtesis,P. *et al.* (2004) PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res.*, **32**, D112–D114.
29. Wu,C.H., Yeh,L.S., 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.
30. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K., Taylor,P. *et al.* (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, **31**, 400–402.
31. Hulo,N., Sigrist,C.J., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, D134–D137.
32. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L.L. (2002) The Pfam protein famililes database. *Nucleic Acids Res.*, **30**, 276–280.
33. Bru,C., Courcelle,E., Carrere,S., Beausse,Y., Dalmar,S. and Kahn,D. (2005) The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res.*, **33**, D212–D215.
34. Servant,F., Bru,C., Carrere,S., Courcelle,E., Gouzy,J., Peyruc,D. and Kahn,D. (2002) ProDom: automated clustering of homologous domains. *Brief Bioinform.*, **3**, 246–251.
35. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.*, **32**, D142–D144.
36. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.
37. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
38. Wootton,J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.*, **18**, 269–285.
39. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
40. Bendtsen,J.D., Nielsen,H., von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol*, **340**, 783–795.
41. Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.*, **8**, 581–599.
42. Sonnhammer,E.L., von Heijne,G. and Krogh,A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 175–182.
43. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome. Res.*, **12**, 1611–1618.
44. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
45. Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.
46. Allen,K.D. (2002) Assaying gene content in *Arabidopsis*. *Proc. Natl Acad. Sci. USA*, **99**, 9568–9572.
47. Shiu,S.H., Karlowski,W.M., Pan,R., Tzeng,Y.H., Mayer,K.F. and Li,W.H. (2004) Comparative analysis of the receptor-like kinase family in *Arabidopsis* and rice. *Plant Cell*, **16**, 1220–1234.

48. Muller,C., Denis,M., Gentzbittel,L. and Faraut,T. (2004) The Iccare web server: an attempt to merge sequence and mapping information for plant and animal species. *Nucleic Acids Res*., **32**, W429–W434.
49. Huan,J., Prins,J., Wang,W. and Vision,T. (2003) Reconstruction of ancestral gene order after segmental duplication and gene loss. In *Proceedings of the Conference on IEEE Computer Society Bioinformatics (CSB)*. IEEE Computer Society, Stanford, CA, USA, pp. 484–485.
50. Simillion,C., Vandepoele,K., Saeys,Y. and Van de Peer,Y. (2004) Building genomic profiles for uncovering segmental homology in the twilight zone. *Genome Res*., **14**, 1095–1106.