BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 2 2004, pages 268–270
DOI: 10.1093/bioinformatics/btg398

The Helmholtz Network for Bioinformatics: an integrative web portal for bioinformatics resources

T. Crass[1,*], I. Antes[2], R. Basekow[3], P. Bork[4,5], C. Buning[6], M. Christensen[1], H. Claußen[6], C. Ebeling[7], P. Ernst[8], V. Gailus-Durner[9], K.-H. Glatting[8], R. Gohla[1], F. Gößling[11], K. Grote[9], K. Heidtke[3], A. Herrmann[5], S. O'Keeffe[9], O. Kießlich[3], S. Kolibal[6], J. O. Korbel[4,5], T. Lengauer[2], I. Liebich[1], M. van der Linden[8], H. Luz[12], K. Meissner[3], C. von Mering[4,5], H.-T. Mevissen[6], H.-W. Mewes[10], H. Michael[1], M. Mokrejs[10], T. Müller[12], H. Pospisil[5], M. Rarey[6], J. G. Reich[5], R. Schneider[9], D. Schomburg[7], S. Schulze-Kremer[3], K. Schwarzer[1], I. Sommer[2], S. Springstubbe[6], S. Suhai[8], G. Thoppae[10], M. Vingron[12], J. Warfsmann[10], T. Werner[9], D. Wetzler[7], E. Wingender[1] and R. Zimmer[6,13]

[1]Department of Bioinformatics, Medical Faculty, Georg August University Göttingen, Goldschmidtstrasse 1, 37077 Göttingen, Germany, [2]Max Planck Institute for Informatics, Computational Biology and Applied Algorithmics, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, [3]RZPD Deutsches Ressourcenzentrum für Genomforschung GmbH, Heubnerweg 6, 14059 Berlin, Germany, [4]European Molecular Biology Laboratory (EMBL), Structural and Computational Biology Programme, Meyerhofstraße 1, 69117 Heidelberg, Germany, [5]Department of Bioinformatics, Max Delbrück Center for Molecular Medicine (MDC), Robert-Roessle-Str. 10, 13125 Berlin, Germany, [6]Fraunhofer Institute for Algorithms and Scientific Computing (formerly GMD), Schloss Birlinghoven, 53754 Sankt Augustin, Germany, [7]Institute of Biochemistry, University of Cologne, Zuelpicher Straße 47, 50674 Cologne, Germany, [8]Department of Molecular Biophysics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany, [9]Institute of Experimental Genetics, [10]Institute for Bioinformatics (MIPS-IBI), National Research Center for Environment and Health (GSF), Ingolstädter Landstraße 1, 85764 Neuherberg/Munich, Germany, [11]Genome Analysis Department, German Biotechnology Research Center (GBF), Mascheroder Weg 1, 38124 Braunschweig, Germany, [12]Bioinformatics Department, Max Planck Institute for Molecular Genetics, Ihnestraße 73, 14195 Berlin, Germany and [13]Institut für Informatik, Praktische Informatik und Bioinformatik, Ludwig Maximilian University Munich, Theresienstraße 39, 80333 Munich, Germany

Received on April 15, 2003; revised on July 24, 2003; accepted on July 31, 2003

ABSTRACT

Summary: The Helmholtz Network for Bioinformatics (HNB) is a joint venture of eleven German bioinformatics research groups that offers convenient access to numerous bioinformatics resources through a single web portal. The '*Guided Solution Finder*' which is available through the HNB portal helps users to locate the appropriate resources to answer their queries by employing a detailed, tree-like questionnaire. Furthermore, automated complex tool cascades ('*tasks*'), involving resources located on different servers, have been implemented, allowing users to perform comprehensive data

*To whom correspondence should be addressed.

analyses without the requirement of further manual intervention for data transfer and re-formatting. Currently, automated cascades for the analysis of regulatory DNA segments as well as for the prediction of protein functional properties are provided.

Availability: The HNB portal is available at http://www.hnbioinfo.de

Contact: torsten.crass@med.uni-goettingen.de

MOTIVATION

The Helmholtz Network for Bioinformatics (HNB) (http://www.hnbioinfo.de/members0306), a joint venture of the Helmholtz Community of Research Centres[†] and other German research institutes,[‡] has been formed to take a step forward from a mere bioinformatics toolbox towards offering web-based, problem-oriented, task-centered solutions that span several bioinformatics tools.

STRATEGY

HNB's offer has three levels.

At the first level (tool box) HNB comprises a wide variety of resources for nucleic acid and protein analysis. Many HNB tools have been preconfigured with standard parameters that were defined by experience and are well suited for the majority of cases. HNB tools include, among others, standard bioinformatics applications like the HUSAR package (Senger *et al*., 1998), the genome analysis tool PEDANT (Frishman *et al*., 2003), SRS (Etzold *et al*., 1996) and the protein function prediction tool STRING (von Mering *et al*., 2003).

At the second level (tool navigation) the HNB portal simplifies the selection of resources for many fundamental bioinformatics tasks, especially for novice users, by offering a so-called 'Guided Solution Finder'. This unique WWW-based interface guides the user through a tree of decision nodes represented by simple questions leading directly to those resources (leaf nodes), that are most appropriate for solving the user's request. This problem-oriented approach does not require any previous knowledge of the available tools and yet allows users to easily identify appropriate solutions for their problems.

At the third level (tool integration), HNB provides mechanisms that allow HNB researchers to integrate programs and databases into automated tool cascades. This task-oriented approach liberates the user from the necessity of manually re-formatting and transferring intermediary result data once a task has been launched.

The user's input and output data are registered via a common API in a central 'virtual user space' and stored on different HNB servers for a defined period of time, allowing users easy access to their own data for re-evaluation and re-use.

We now describe integrated tool cascades currently offered by HNB.

Genomic sequence analysis

Integrated automated tool cascades for predicting and annotating putative regulatory regions in eukaryotic genomes have been realized using resources developed at GBF, GSF and partly in co-operation with commercial partners.[§]

Within a cascade called *TF Scan*, a nucleotide sequence is simultaneously submitted to the transcription factor (TF)-binding site prediction programs *PatSearch* (Wingender *et al*., 1997) and *MatInspector* (Quandt *et al*., 1995), and their output is subsequently combined into a single result page, with all site hits being linked to the TRANSFAC (Matys *et al*., 2003) and EMBL (Stoesser *et al*., 2002) databases. The more complex *RegRegion Analysis* cascade first runs *PromoterInspector* (Scherf *et al*., 2000) to scan a nucleotide sequence for putative promoter regions and subsequently calls *TF Scan* on each identified candidate promoter. Finally, a *Genomic Mapping* task maps a query sequence against various eukaryotic genomes and provides TF-binding site annotation of the identified genomic regions by extending the *Ensembl!* suite (Hubbard *et al*., 2002).

Protein sequence analysis

The focus of this HNB-subsection is on the prediction of protein features, taking into account the close relationship between protein structure predictions, protein family analysis and protein function prediction. Protein family analysis is performed by searching against the SYSTERS cluster set (Krause *et al*., 2002), a hierarchical classification of all SWISS-PROT (Bairoch and Apweiler, 2000), TrEMBL and PIR (Barker *et al*., 2001) sequences into disjoint protein family clusters and superfamilies. Protein function prediction focuses on domain structure prediction using SMART (Letunic *et al*., 2002). Protein structure prediction is performed using a threading algorithm (Alexandrov *et al*., 1996; Zien *et al*., 2000) comparing a query sequence with a representative subset of all protein structures from PDB (Berman *et al*., 2000). At the time of print of this note we expect the STRING Server (von Mering *et al*., 2003) to be integrated into the protein analysis task, as well.

By submitting a query sequence to a general *Protein Analysis* task, a combined summary of the results from all of

[†]Four centres thereof in particular: German Biotechnology Research Center (GBF), Braunschweig; German Cancer Research Center (DKFZ), Heidelberg; Max Delbrück Center for Molecular Medicine (MDC), Berlin-Buch; National Research Centre for Environment and Health (GSF), Neuherberg/Munich.

[‡]Department of Bioinformatics, Medical Faculty, Georg August University Göttingen; Fraunhofer Institute for Algorithms and Scientific Computing (formerly GMD), St. Augustin; Institute of Biochemistry, University of Cologne; Max Planck Institute for Informatics, Saarbrücken; Max Planck Institute for Molecular Genetics, Berlin; RZPD Deutsches Ressourcenzentrum für Genomforschung GmbH, Berlin.

[§]BIOBASE GmbH, Biomax Informatics AG and Genomatix Software GmbH.

the above-mentioned tools is generated, which also provides links to the individual intermediate results of the stand-alone tools. These tools can then be re-entered for the refinement of the overall analysis.

TECHNICAL CONSIDERATIONS

HNB is based on a heterogeneous network of servers, distributed over different Internet domains, some of which are protected by firewalls. To overcome the resulting restrictions on inter-server communication, the HTTP/HTTPS is used as the transport-tunnelling layer for the actual data exchange via XML-based communication protocols (including SOAP).

Although anonymous access is possible for many HNB-resources, a certificate-based HNB-user authentication mechanism had to be implemented to accommodate the restricted user access (i.e. 'academic only') to certain resources. User certificates can easily be obtained through online registration. HNB is free for academic users; restrictions apply only for commercial users.

ACKNOWLEDGEMENTS

This work has been funded by a grant from the Federal Ministry of Education and Research (01SF9988/4).

REFERENCES

Alexandrov,N., Nussinov,R. and Zimmer,R. (1996) Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. In Hunter,L. and Klein,T.E. (eds), *Pacific Symposium on Biocomputing'96*. World Scientific Publishing Co. Pte. Ltd., Singapore, pp. 53–69.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL. *Nucleic Acids Res.*, **28**, 45–48.

Barker,W.C., Garavelli,J.S., Hou,Z., Huang,H., Ledley,R.S., McGarvey,P.B., Mewes,H.W., Orcutt,B.C., Pfeiffer,F., Tsugita,A. *et al*. (2001) Protein information resource: a community resource for expert annotation of protein data. *Nucleic Acids Res.*, **29**, 29–32.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.

Frishman,D., Mokrejs,M., Kosykh,D., Karstenmuller,G., Kolesov,G., Zubrzycki,I., Gruber,C., Geier,B., Kaps,A., Volz,A., Wagner,C. *et al*. (2003). The Pedant genome database. *Nucleic Acids Res.*, **31**, 207–211.

Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,I., Cox,T., Cuff,J., Curwen,V., Down,T. *et al*. (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.

Krause,A., Haas,S.A., Coward,E. and Vingron,M. (2002) SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. *Nucleic Acids Res.*, **30**, 299–300.

Letunic,I., Goodstadt,L., Dickens,N.J., Doerks,T., Schultz,J., Mott,R., Ciccarelli,F., Copley,R.R., Ponting,C.P. and Bork,P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.*, **30**, 242–244.

Matys,V., Fricke,E., Geffers,R., Gößling,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.

Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector—new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.

Scherf,M., Klingenhoff,A. and Werner,T. (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context-sensitive approach. *J. Mol. Biol.*, **297**, 599–606.

Senger,M., Flores,T., Glatting,K., Ernst,P., Hotz-Wagenblatt,A. and Suhai,S. (1998) W2H: WWW interface to the GCG sequence analysis package. *Bioinformatics*, **14**, 452–457.

Stoesser,G., Baker,W., van den Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V. *et al*. (2002) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **30**, 21–26.

von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.

Wingender,E., Karas,H. and Knüppel,R. (1997) TRANSFAC database as a bridge between sequence data libraries and biological function. In Altmann,R.B., Dunker,A.K., Hunter,L. and Klein,T.E. (eds), *Pacific Symposium on Biocomputing'97*. World Scientific Publishing Co. Pte. Ltd., Singapore, pp. 477–485.

Zien,A., Zimmer,R. and Lengauer,T. (2000) A simple iterative approach to parameter optimization. *J. Comput. Biol.*, **7**, 483–501.