Sentra, a database of signal transduction proteins

Natalia Maltsev*, E. Marland, G. X. Yu, S. Bhatnagar and R. Lusk

Mathematics and Computer Science Division, Argonne National Laboratory, 9700 S. Cass Avenue, Argonne, IL 60439, USA

Received September 18, 2001; Accepted October 10, 2001

ABSTRACT

Sentra (http://www-wit.mcs.anl.gov/sentra) is a database of signal transduction proteins with the emphasis on microbial signal transduction. The database was updated to include classes of signal transduction systems modulated by either phosphorylation or methylation reactions such as PAS proteins and serine/threonine kinases, as well as the classical two-component histidine kinases and methyl-accepting chemotaxis proteins. Currently, Sentra contains signal transduction proteins from 43 completely sequenced prokaryotic genomes as well as sequences from SWISS-PROT and TrEMBL. Signal transduction proteins are annotated with information describing conserved domains, paralogous and orthologous sequences, and conserved chromosomal gene clusters. The newly developed user interface supports flexible search capabilities and extensive visualization of the data.

INTRODUCTION

A prokaryotic organism's ability to monitor its extracellular and intracellular environment is vital for its adaptation and survival. Complex signaling and regulatory networks govern transcriptional responses to various environmental and developmental conditions. A detailed characterization of proteins that constitute such networks is essential for understanding the signal transduction process. Prokaryotes use a variety of mechanisms to sense and transmit information that is important for maintaining homeostatic balance. Most of the microbial signaling proteins are assembled from modular components that include a variety of input sensing domains, output effector domains and domains for promoting protein–protein communications (1). Understanding the sensory process requires knowledge of the nature of the transmitted signal as well as mechanisms involved in its transduction.

The Sentra database (http://www-wit.mcs.anl.gov/sentra) was designed to allow identification and detailed computational analysis of proteins involved in signal transduction using comprehensive sequence analysis tools. These tools enable the user to analyze conserved functional domains, to identify paralogous and orthologous sequences as well as conserved chromosomal gene clusters. The objective was to add sensitivity to the sequence analysis of signal transduction proteins and to assist researchers in predicting the nature of a transmitted signal and possible mechanisms of its transduction.

DESCRIPTION OF THE DATABASE

Identification of signal transduction proteins in Sentra

The current version of Sentra has been updated to include other classes of signal transduction systems that are modulated by phosphorylation or methylation reactions, such as PAS proteins, serine/threonine kinases and phosphatases, besides the classical two-component signal transduction histidine kinases and methyl-accepting chemotaxis proteins. Sentra now contains signal transduction proteins from 43 completely sequenced prokaryotic genomes, as well as sequences from public databases (SWISS-PROT and TrEMBL) (2). We have identified and deposited in Sentra proteins relevant to the signal transduction process. During the past year we have significantly modified the data annotation process in Sentra and its database architecture in order to improve performance and simplify the process of updates.

Identification and analysis of signal transduction proteins in Sentra database includes the following steps. First, the sequences from 43 completely sequenced genomes plus the sequences from the SWISS-PROT and TrEMBL databases were analyzed by using the domain analysis tools from Pfam (3) and Blocks (4,5). Then, the results of these analyses were parsed and stored in a relational (Oracle) database. A subset of putative signal transduction proteins, containing conserved domains relevant to signal transduction [e.g. histidine kinase domains (PF00512, PF02518), the histidine-receiving domain (PF00072)] with the E-value not more than 1.0e–03 was extracted and used for further analysis and annotation in Sentra.

Analysis of proteins in Sentra

Large-scale homology searches with BLAST (6) against the non-redundant database (7) and sequenced genomes were used in order to assist in assigning functions to putative signal transduction proteins. We used the SOSUI program (8) to help identify receptors transmitting environmental signals, and we used the method described by Overbeek *et al.* (9) to analyze conserved chromosomal gene clusters containing signal transduction proteins. The latter technique attempts to infer functional coupling between genes based on conservation of chromosomal gene clusters between genomes. Our analysis showed that a significant number of signal transduction proteins occur as part of a sensory transduction cassette containing genes that encode transmitters and receivers. In some cases, however, signal transduction genes form conserved clusters with genes encoding the effectors (e.g. chemotaxis operon, ABC transporters). The composition of a chromosomal gene cluster can provide important clues about the probable

*To whom correspondence should be addressed. Tel: +1 630 252 5195; Fax: +1 630 252 5986; Email: maltsev@mcs.anl.gov nature of the transmitted signal and the composition of the regulatory cascade. All of the proteins deposited in Sentra are annotated with the information describing functional domains that co-occur with the signal transduction domains, such as PAS domains and sigma54 domains. Such analysis of the domain composition provides significant information about the mechanisms involved in signal processing by a particular protein.

Representation of the data in Sentra

A number of additional tools for accessing and representing the data in Sentra were developed in the past year. These tools include SQL-based searching capabilities that allow the user to execute flexible queries and to search Sentra by organism, by domain profile, by keyword and by protein identification. A toolkit for visualizing the genome maps, conserved chromosomal gene clusters and domain composition was added to Sentra. Additionally, Sentra now supports comparative and evolutionary analysis of signal transduction proteins. The database also provides links to the WIT2 (10) environment and a number of public databases (e.g. NCBI, SWISS-PROT, Pfam, Blocks).

FUTURE PROSPECTS

We plan to further update Sentra with newly sequenced prokaryotic genomes and to include expansion of Sentra contents to include graphical representation of prokaryotic signal transduction pathways and allow the user to annotate the data. Another planned development is the design of ontology for prokaryotic signal transduction genes and allowing the distribution of Sentra content in XML format.

SUPPLEMENTARY MATERIAL

The following information is available as Supplementary Material at NAR Online: a statistical table showing the number of histidine kinase and methyl-accepting chemotaxis proteins detected within each of the 43 complete genomes; a statistical table showing the number of certain domains found in proteins detected per organism; a flow chart describing the process involved in creating Sentra; representative examples of analysis in Sentra.

ACKNOWLEDGEMENT

This work was supported by the Office of Biological and Environmental Research, US Department of Energy, under Contract W-31-109-Eng-38.

REFERENCES

1. Parkinson,J.S. (1995) In Hoch,A. and Silhavy,T.J. (eds), *Two-Component Signal Transduction.* ASM Press, Washington, DC, p. 923.
2. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
3. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe K.L. and Sonnhammer,E.L.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 276–280.
4. Henikoff,J.G., Greene E.A., Pietrokovski,S. and Henikoff,S. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.*, **28**, 228–230.
5. Henikoff,S., Henikoff,J.G. and Pietrokovski,S. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, **15**, 471–479.
6. 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.
7. Benson,D.A., Boguski,M.S., Lipman,D.J, Ostell,J. and Ouellette,B.F.F. (1998) GenBank. *Nucleic Acids Res.*, **26**, 1–6. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 17–20.
8. Hirokawa,T., Boon-Chieng,S. and Mitaku,S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379.
9. Overbeek,R., Fonstein,M., DSouza,M., Pusch,G.D. and Maltsev,N. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
10. Overbeek,R., Larsen,N., Maltsev,N., Pusch,G.D. and Selkov,E. (1999) In Letovsky,S. (ed.), *Molecular Biology Databases*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 158–163.
11. Taylor,B.L. and Zhulin,I.B. (1999) PAS domains: internal sensors of oxygen, redox potential, and light. *Microbiol. Mol. Biol. Rev.*, **63**, 479–506.
12. Aravind,L. and Ponting,C.P. (1999) The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signaling proteins. *FEMS Microbiol. Lett.*, **176**, 111–116.