*Published online 29 November 2006* *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D271–D273***
*doi:10.1093/nar/gkl949*

# Sentra: a database of signal transduction proteins for comparative genome analysis

**Mark D'Souza[1,2,*], Elizabeth M. Glass[1,2], Mustafa H. Syed[1], Yi Zhang[1], Alexis Rodriguez[1], Natalia Maltsev[1,2] and Michael Y. Galperin[3,*]**

[1]Computational Biology Group, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA, [2]Computation Institute, University of Chicago, Chicago, IL 60637, USA and [3]National Center for Biotechnology Information, National Library of Medicine, MSC3830, National Institutes of Health, Bethesda, MD 20894, USA

Received September 14, 2006; Revised October 18, 2006; Accepted October 20, 2006

## ABSTRACT

**Sentra (http://compbio.mcs.anl.gov/sentra), a database of signal transduction proteins encoded in completely sequenced prokaryotic genomes, has been updated to reflect recent advances in understanding signal transduction events on a whole-genome scale. Sentra consists of two principal components, a manually curated list of signal transduction proteins in 202 completely sequenced prokaryotic genomes and an automatically generated listing of predicted signaling proteins in 235 sequenced genomes that are awaiting manual curation. In addition to two-component histidine kinases and response regulators, the database now lists manually curated Ser/Thr/Tyr protein kinases and protein phosphatases, as well as adenylate and diguanylate cyclases and c-di-GMP phosphodiesterases, as defined in several recent reviews. All entries in Sentra are extensively annotated with relevant information from public databases (e.g. UniProt, KEGG, PDB and NCBI). Sentra's infrastructure was redesigned to support interactive cross-genome comparisons of signal transduction capabilities of prokaryotic organisms from a taxonomic and phenotypic perspective and in the framework of signal transduction pathways from KEGG. Sentra leverages the PUMA2 system to support interactive analysis and annotation of signal transduction proteins by the users.**

## INTRODUCTION

Recent experimental and *in silico* studies have resulted in a much better understanding of the principles and mechanisms of prokaryotic signal transduction (1–6). The list of recognized environmental sensors has been dramatically expanded and now includes, in addition to two-component histidine kinases and methyl-accepting chemotaxis proteins, Ser/Thr/Tyr protein kinases and protein phosphatases, as well as adenylate and diguanylate cyclases and c-di-GMP phosphodiesterases (2–10). These classes of proteins are also found as (predicted) cytoplasmic proteins, proposed to function as sensors of the intracellular biochemical parameters, such as pH, osmolarity or levels of oxygen, CO, NO and other molecules (2,10). Accordingly, many prokaryotic genomes contain multiple copies of the respective genes, whose exact functions (i.e. the parameters sensed by their protein products) are rarely known. Detailed analyses of protein sets involved in signal transduction in such model organisms as *Escherichia coli, Bacillus subtilis, Pseudomonas aeruginosa, Synechocystis* sp. PCC6803, *Anabaena* sp. PCC7120 or *Halobacterium salinarum* brought very interesting results and provided needed insight into the signal transduction mechanisms. *In silico* studies have contributed by highlighting such phenomena as the abundance of (predicted) diguanylate cyclases and c-di-GMP phosphodiesterases in many bacterial genomes, the importance of cross-talk between different signaling pathways and the existence of a complex system of intracellular signaling (2,3,10).

Progress in understanding of prokaryotic signal transduction systems, as well as availability of a large number of newly sequenced genomes, prompted us to perform a major update of Sentra (http://compbio.mcs.anl.gov/sentra), a database of signal transduction proteins developed by the Bioinformatics group at Argonne National Laboratory (13,14). The objective of further development of Sentra was to provide users with an analytical environment containing expert-curated information describing prokaryotic signal transduction systems, as well as up-to-date knowledge base and interactive analytical tools for further analysis of signal transduction proteins in all completely sequenced genomes as they become publicly available. Such an environment will add accuracy and sensitivity to the sequence analysis of

*To whom correspondence should be addressed. Tel: +1 630 252 5195; Fax: +1 630 252 5986; Email: dsouza@mcs.anl.gov
*Correspondence may also be addressed to Michael Y. Galperin. Tel: +1 301 435 5910; Fax: +1 301 435 7793; Email: galperin@ncbi.nlm.nih.gov

signal transduction proteins and aid in the development of conjectures regarding the nature of the transmitted signal. The previous release of Sentra featured signal transduction proteins encoded in 43 completely sequenced genomes (14). Although it contained all complete, public genomes at the time of publication, it was missing a number of valuable data and analytical capabilities. For example, it did not include diguanylate cyclases or c-di-GMP phosphodiesterases and did not support cross-genome comparative analysis of signal transduction systems (14). Further, since most components of the signal transduction machinery are multi-domain proteins, they are notoriously difficult to annotate through automated sequence comparisons and are commonly misannotated in genomic databases (10,15). Discovery of new domains often makes the existing annotations incomplete or even obsolete. To provide the solution to this problem, Sentra was redesigned to perform periodic (monthly) automated updates that include automated pre-computed analysis of newly sequenced genomes and re-analysis of existing Sentra genomes with an array of bioinformatics tools including InterPro (16), Blocks (17), BLAST (18), TMHMM (19) and tools developed by our group (e.g. Dremmel, http://compbio.mcs.anl.gov/dremmel and Chisel, http://compbio.mcs.anl.gov/CHISEL). The results of these automated analyses are presented to the users in Sentra's interactive environment for further updates and annotation. The most significant changes in Sentra database content, capabilities and user interface are as follows.

### Update of the Sentra database content

Sentra now consists of two principal components: (i) a manually curated list of signal transduction proteins that includes proteins derived from 202 completely sequenced prokaryotic genomes, and (ii) an automatically generated listing of predicted signaling proteins in 235 genomes that are awaiting manual curation.

The expert-curated section of the database now lists, besides two-component histidine kinases and response regulators, Ser/Thr/Tyr protein kinases and protein phosphatases, as well as adenylate and diguanylate cyclases and c-di-GMP phosphodiesterases, as defined in several recent reviews (2,10,12).

### Support for comparative and evolutionary analysis of signal transduction proteins and signaling pathways

In the process of adaptation to environment, prokaryotic organisms have developed an ability to detect and process environmental signals that are vital for their survival. Sentra provides a unique opportunity to explore and compare the signaling apparatus of prokaryotes according to their habitat (e.g. aquatic, terrestrial), lifestyle (e.g. pathogenic) and major physiological features (e.g. energy source, motility). Users can also perform comparative analysis of signal transduction proteins characteristic of different taxonomic groups of organisms in the framework of the signaling pathways from the KEGG database (20). This capability allows identification of signaling pathways and mechanisms characteristic of particular taxonomic groups and habitats.

Sentra leverages the PUMA2 (21) system for high-throughput analysis of genomes being developed by the Bioinformatics group at Argonne. Such a connection allows Sentra to support comparative analysis of the prokaryotic signal transduction systems at multiple levels of organization: users may explore domain and feature composition of signal transduction proteins and perform interactive analysis of sequences by over 30 bioinformatics tools. All entries in Sentra are annotated with the information from the PUMA2 knowledge base integrating information from over 20 sequence, structural, metabolic and taxonomic databases, as well as the derived results from various bioinformatics tools. Sentra also contains information regarding participation of the signal transduction proteins in conserved chromosomal gene clusters (22). Such information may provide important clues regarding the nature of the transmitted signal.

### Support for user annotation of signal transduction proteins

One of the important new features of Sentra is its support for the user annotation of the signal transduction proteins via the PUMA2 framework. Registered users can interactively analyze the sequences, correct functional assignment and provide detailed comments. Such capability will allow us to leverage an enormous expert knowledge accumulated in the scientific community for annotation of information in the Sentra database. All computationally intensive operations in Sentra are performed using the Grid technology-based engine GADU (23) being developed by the Bioinformatics group at Argonne.

### Future prospects

As new completely sequenced microbial genomes become publicly available, they will be processed through the automated pipeline and included in quarterly updates of the database. These genomes will also be subject to manual curation of the overall protein lists and orthology groupings. We also intend to provide manually curated lists of proteins containing certain signal transduction domains, such as PAS (24) and FHA (25).

## ACKNOWLEDGEMENTS

This work was supported by the Office of Biological and Environmental Research, US Department of Energy, under Contract DE-AC02-06CH11357 and by the Intramural Research Program of the NIH, National Library of Medicine. N.M. and E.M.G. acknowledge membership within and support in part from the Region V 'Great Lakes' Regional Center of Excellence in Biodefense and Emerging Infectious Diseases Consortium (GLRCE, NIAID Award 1-U54-AI-057153). M.D. acknowledges membership and support to NMPDR Bioinformatics Resource Center NIH/NIAID (Award NNSN 266200400042C). We are grateful to Luke Ulrich for his work on PhyloBlocks. Funding to pay the Open Access publication charges for this article was provided by the Intramural Research Program of the NIH, National Library of Medicine.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Inouye,M. and Dutta,R. (eds). (2003) *Histidine kinases in signal transduction*. Academic Press, San Diego, London.

2. Galperin,M.Y. (2004) Bacterial signal transduction network in a genomic perspective. *Environ. Microbiol.*, **6**, 552–567.
3. Ulrich,L.E., Koonin,E.V. and Zhulin,I.B. (2005) One-component systems dominate signal transduction in prokaryotes. *Trends Microbiol.*, **13**, 52–56.
4. Galperin,M.Y. (2006) Structural classification of bacterial response regulators: diversity of output domains and domain combinations. *J. Bacteriol.*, **188**, 4169–4182.
5. Ashby,M.K. (2006) Distribution, structure and diversity of 'bacterial' genes encoding two-component proteins in the Euryarchaeota. *Archaea*, **2**, 11–30.
6. Zhang,W. and Shi,L. (2005) Distribution and evolution of multiple-step phosphorelay in prokaryotes: lateral domain recruitment involved in the formation of hybrid-type histidine kinases. *Microbiology*, **151**, 2159–2173.
7. Terauchi,K. and Ohmori,M. (2004) Blue light stimulates cyanobacterial motility via a cAMP signal transduction system. *Mol. Microbiol.*, **52**, 303–309.
8. Jenal,U. (2004) Cyclic di-guanosine-monophosphate comes of age: a novel secondary messenger involved in modulating cell surface structures in bacteria? *Curr. Opin. Microbiol.*, **7**, 185–191.
9. Römling,U., Gomelsky,M. and Galperin,M.Y. (2005) C-di-GMP: The dawning of a novel bacterial signaling system. *Mol. Microbiol.*, **57**, 629–639.
10. Galperin,M.Y. (2005) A census of membrane-bound and intracellular signal transduction proteins in bacteria: bacterial IQ, extroverts and introverts. *BMC Microbiol.*, **5**, 35.
11. Shenroy,A.R. and Visweswariah,S.S. (2004) Class III nucleotide cyclases in bacteria and archaebacteria: lineage-specific expansion of adenylyl cyclases and a dearth of guanylyl cyclases. *FEBS Lett.*, **561**, 11–21.
12. Krupa,A. and Srinivasan,N. (2005) Diversity in domain architectures of Ser/Thr kinases and their homologues in prokaryotes. *BMC Genomics*, **6**, 129.
13. D'Souza,M., Romine,M.F. and Maltsev,N. (2000) SENTRA, a database of signal transduction proteins. *Nucleic Acids Res.*, **28**, 335–336.
14. Maltsev,N., Marland,E., Yu,G.X., Bhatnagar,S. and Lusk,R. (2002) Sentra, a database of signal transduction proteins. *Nucleic Acids Res.*, **30**, 349–350.
15. Zhulin,I.B., Nikolskaya,A.N. and Galperin,M.Y. (2003) Common extracellular sensory domains in transmembrane receptors for diverse signal transduction pathways in bacteria and archaea. *J. Bacteriol.*, **185**, 285–294.
16. 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.
17. 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.
18. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zheng,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.
19. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
20. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
21. Maltsev,N., Glass,E., Sulakhe,D., Rodriguez,A., Syed,M.H., Bompada,T., Zhang,Y. and D'Souza,M. (2006) PUMA2—grid-based high-throughput analysis of genomes and metabolic pathways. *Nucleic Acids Res.*, **34**, D369–D372.
22. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
23. Sulakhe,D., Rodriguez,A., D'Souza,M., Wilde,M., Nefedova,V., Foster,I. and Maltsev,N. (2005) GNARE: automated system for high-throughput genome analysis with grid computational backend. *J. Clin. Monit. Comput.*, **19**, 361–369.
24. 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.
25. Pallen,M., Chaudhuri,R. and Khan,A. (2002) Bacterial FHA domains: neglected players in the phospho-threonine signaling game? *Trends Microbiol.*, **10**, 556–563.