 Nucleic Acids Research, 2000, Vol. 28, No. 1 **123–125**

WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction

Ross Overbeek[1,*], Niels Larsen[1], Gordon D. Pusch[1], Mark D'Souza[1,2], Evgeni Selkov Jr[1,2], Nikos Kyrpides[1], Michael Fonstein[1], Natalia Maltsev[2] and Evgeni Selkov[1,2]

[1]Integrated Genomics Inc., 2201 W. Campbell Park Drive, Chicago, IL 60612, USA and [2]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA

Received September 1, 1999; Revised and Accepted October 13, 1999

ABSTRACT

The WIT (What Is There) (http://wit.mcs.anl.gov/WIT2/) system has been designed to support comparative analysis of sequenced genomes and to generate metabolic reconstructions based on chromosomal sequences and metabolic modules from the EMP/MPW family of databases. This system contains data derived from about 40 completed or nearly completed genomes. Sequence homologies, various ORF-clustering algorithms, relative gene positions on the chromosome and placement of gene products in metabolic pathways (metabolic reconstruction) can be used for the assignment of gene functions and for development of overviews of genomes within WIT. The integration of a large number of phylogenetically diverse genomes in WIT facilitates the understanding of the physiology of different organisms.

INTRODUCTION

Starting with *Haemophilus influenza* (1) in 1995, over 20 microbial organisms have had their total genomic DNA sequenced and almost 100 others have been started as shown in the GOLD database (2). Currently we are observing an impressive development of the human genome project (3,4). In response to this growing amount of sequence data, computational tools for genome analysis have been developed and merged into shared analytical environments, such as GeneQuiz (5), KEGG (6), Pedant (7) and Entrez Genomes (8), moving cross-genome analysis to a new level. The development of analytical systems, together with the growth of sequencing data, have increased gene recognition rates from <50% (9,10) to >70% (11,12). Today, this remaining 30%, so-called 'hypothetical' or 'orphan' genes, separates us from a complete description of the genomic content and functions of an organism.

Computational approaches based on various types of clustering of potential genes, whether in phylogenetic space, as clusters of orthologous genes (COGs) (13) or position on the chromosome, such as in operons (14), increase the gene assignment level even further. An important stage of genome analysis is the integration of gene assignments into an organism-specific overview via so-called functional reconstruction (15), which is the conceptual assembly of metabolic pathways, transport units and signal transduction pathways. It allows reconciliation of inconsistencies between different types of analysis, and often results in changes of initial gene function assignments based on similarity scoring.

The WIT system, discussed in this paper, represents the development of a genome analysis strategy in a multi-genome environment, which combines a variety of tools, dealing with individual open reading frames (ORFs) or proteins, with the ability to derive general conclusions. Using the WIT genome analysis system, a major part of the central metabolism of an organism can be reconstructed entirely *in silico* (16).

WIT: A VIEW TO A GENOME

The current version of the WIT system is available at Argonne National Laboratory (http://wit.mcs.anl.gov/WIT2) or at Integrated Genomics Inc. (http://wit.IntegratedGenomics.com/IGwit) and contains 43 complete or nearly complete genomes (Table 1).

These genomes consist of 123 482 predicted ORFs, of which 78 144 could be given functional assignments and 41 742 could be assembled into metabolic pathways, which came from EMP/MPW database (15). Pathways involved in the metabolism of carbohydrates and amino acids are connected into schematic overviews allowing the user to reveal substrates and final products connecting metabolic modules.

In order to incorporate a genome into WIT, a gene-searching program called CRITICA (17) can be used. Potential coding regions recognized in the DNA contigs are subjected to a FASTA search against the non-redundant database of assigned genes and loaded into the WIT system, together with the pre-computed tables of best hits.

WIT provides a set of tools for the characterization of gene structures and functions, such as Functional Coupling, or Preserved Operons. WIT also provides integrated WWW access to such tools as PSI-BLAST, PROSITE, ProDom, COG, ClustalW and others. Functional content may be queried, for example, by looking for specific functions missing in the metabolic pathways, or by separating alternative gene functions derived from similarities found for a putative gene.

After genes have been assigned initial functions, they are then 'attached' to pathways by choosing templates from metabolic database (MPW) which best incorporate all observed functions. For any given organism, this usually leads to identification of

*To whom correspondence should be addressed. Tel: +1 312 491 0846; Fax: +1 312 491 0856; Email: ross@integratedgenomics.com

Table 1. Genomes in WIT

Eukarya	*Saccharomyces cerevisiae, Caenorhabditis elegans*
Archaea	*Sulfolobus solfataricus, Archaeoglobus fulgidus,Halobacterium* sp., *M.thermoautotrophicum, M.jannaschii, Pyrococcus furiosus, Pyrococcus horikoshii*
Bacteria	*A.aeolicus, C.trachomatis, Synechocystis* sp., *P.gingivalis, M.leprae, M.tuberculosis, B.subtilis, C.acetobutylicum, E.faecalis, M.genitalium, M.pneumoniae, S.pneumoniae, S.pyogenes, Rhizobium* sp., *R.capsulatus, S.aromaticivorans, N.gonorrhoeae, N.meningitidis, C.jejuni, H.pylori, E.coli, Y.pestis, H.influenzae, P.aeruginosa, B.burgdorferi, T.pallidum, D.radiodurans*
Additional Genomes on the public server at Integrated Genomics Inc.	*A.pernix, M.bovis, C.tepidum, S.typhi, T.maritima, A.actinomycetemcomitans, E.nidulans, Oryza sativa, A thaliana, R.prowazekii, P.abysii, C.pneumoniae, C.reinhardtii*

functional sub-systems, as a model for further refinement. For example, it is now possible to identify inconsistencies, potentially missing enzymes/ORFs, thereby refining the model. When a basic model has been created, a curator finally evaluates this model against biochemical data and phenotypes known from the literature. The models come in both textual and graphical representations, fully linked with all underlying data. We call this whole process metabolic reconstruction, and the main role of the WIT system is to support this effort.

To examine or curate a functional model of an organism, one can use functions such as: Compare assignments, Summary of asserted functions and pathways, Examine trimmed ortholog clusters, Examine COG/trimmed ortholog cluster relationships, Search for pathways by regular expression, Search ORF functions by regular expression, Search ORF sequences by similarity search, Find NCBI's MEDLINE-references by EC-number, Search EMP by EC-number, and Find common proteins for organisms. Chromosomal clustering of functionally related genes (14) is another powerful component of the system, which recently allowed us to propose a number of candidate ORFs for 'orphan' metabolic functions. Continuous integration of newly sequenced genomes increases the depth of functional description by a reiterative process.

GAPPED GENOMES IN WIT

An important feature of the WIT system is its emphasis on incomplete or gapped genomes. Algorithms used for gene assignments depend on the size of a dataset used to cluster properties of ORFs, whether it is chromosomal position or ortholog clustering based on bi-directional best hits. By the incorporation of gapped genomes, even the public version of WIT has integrated twice as much data as can be collected from only the completed genomes.

We believe that integrating systems like WIT can offer a solution for the problem of efficient use of incomplete sequence data. The gapped sequence contains a piece of almost every ORF, which allows the assignment of functions to almost all ORFs and the accurate reconstruction of the metabolism of the organism; good informatics can compensate for poorer sequence quality. A comparison of the results of analysis of the gapped genome of *Pseudomonas aeruginosa* with the complete genomes of *Escherichia coli* and *Bacillus subtilis* proves this statement (Table 2).

CONCLUSIONS

WIT has been designed to extract functional content from genome sequences and organize it into a coherent system, in order to facilitate post-sequencing experimental biology. The WIT system provides a set of local tools, which can be used to investigate functions of individual ORFs, based on similarities, motifs and various types of ORF clustering. It also generates overviews of functional subsystems and means to connect them into a complete picture of cellular functionality.

The WIT system is undergoing constant improvements, which can be traced in the PUMA–WIT–WIT2 line of development, and we believe that numerous further additions are needed to provide an adequate toolbox for the biological research community. Major directions of the ongoing WIT development are the following: (i) integration of structural data, which are currently underutilized in WIT; (ii) further development of the collection of functional maps and construction of more abstract scalable overviews, which should eventually cover all cellular functionality, and; (iii) development of a framework, which

Table 2. Comparison of the gapped *P.aeruginosa* genome with those of *E.coli* K-12 and *B.subtilis* 168

	Pseudomonas aeruginosa	*Escherichia coli*	*Bacillus subtilis*
Genome Size (Mb)	6.2	4.7	4.1
DNA assembled (%)	99	100	100
Total ORFs	5627	4289	4083
Assigned ORFs	4191	3499	3016
Asserted pathways	581	906	782
Missing assignments	133	102	178
No sequences	115	233	173

will integrate a flood of the differential display expression array data into the metabolic context.

ACKNOWLEDGEMENTS

Development of WIT was supported in part by the Office of Biological and Environmental Research, US Department of Energy, under Contract W-31-109-Eng-38, and in part by Integrated Genomics Inc.

REFERENCES

1. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A. and Merrick,J.M. (1995) *Science*, **269**, 496–512.
2. Kyrpides,N.C. (1999) *Bioinformatics*, **15**, 773–774.
3. Collins,F.S., Patrinos,A., Jordan,E., Chakravarti,A., Gesteland,R. and Walters,L. (1998) *Science*, **282**, 682–689.
4. Venter,J.C., Adams,M.D., Sutton,G.G., Kerlavage,A.R., Smith,H.O. and Hunkapiller,M. (1998) *Science*, **280**, 1540–1542.
5. Andrade,M.A., Brown,N.P., Leroy,C., Hoersch,S., de Daruvar,A., Reich,C., Franchini,A., Tamames,J., Valencia,A., Ouzounis,C. and Sander,C. (1999) *Bioinformatics*, **15**, 391–412.
6. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) *Nucleic Acids Res.*, **27**, 29–34.
7. Mewes,H.W., Heumann,K., Kaps,A., Mayer,K., Pfeiffer,F., Stocker,S. and Frishman,D. (1999) *Nucleic Acids Res.*, **27**, 44–48. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 37–40.
8. Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouellette,B.F., Rapp,B.A. and Wheeler,D.L. (1999) *Nucleic Acids Res.*, **27**, 12–7. Updated article in this issue: *Nucleic Acids Res*. (2000), **28**, 15–18.
9. Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., FitzGerald,L.M., Clayton,R.A., Gocayne,J.D., Kerlavage,A.R., Dougherty,B.A., Tomb,J.F., Adams,M.D., Reich,C.I., Overbeek,R., Kirkness,E.F., Weinstock,K.G., Merrick,J.M., Glodek,A., Scott,J.L., Geoghagen,N. and Venter,J.C. (1996) *Science*, **273**, 1058–1073.
10. Kaneko,T., Sato,S., Kotani,H., Tanaka,A., Asamizu,E., Nakamura,Y., Miyajima,N., Hirosawa,M., Sugiura,M., Sasamoto,S., Kimura,T., Hosouchi,T., Matsuno,A., Muraki,A., Nakazaki,N., Naruo,K., Okumura,S., Shimpo,S., Takeuchi,C., Wada,T., Watanabe,A., Yamada,M., Yasuda,M. and Tabata,S. (1996) *DNA Res.*, **3**, 185–209.
11. Vlcek,C., Paces,V., Maltsev,N., Haselkorn,R. and Fonstein,M. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 9384–9388.
12. Selkov,E., Maltsev,N., Olsen,G.J., Overbeek,R. and Whitman,W.B. (1997) *Gene*, **197**, GC11–GC26.
13. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) *Science*, **278**, 631–637.
14. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) *Proc. Natl Acad. Sci, USA*, **96**, 2896–2901.
15. Selkov,E., Basmanova,S., Gaasterland,T., Goryanin,I., Gretchkin,Y., Maltsev,N., Nenashev,V., Overbeek,R., Panyushkina,E., Pronevitch,L., Selkov,E.,Jr and Yunus,I. (1996) *Nucleic Acids Res.*, **24**, 26–28.
16. Selkov,E., Overbeek,R., Kogan,Y., Chu,L., Vonstein,V., Holmes,D., Silver,S., Haselkorn,R. and Fonstein,M. (1999) *Proc. Natl. Acad. Sci. USA.*, in press.
17. Badger,J.H. and Olsen,G.J. (1999) *Mol. Biol. Evol.*, **16**, 512–524.