CHAPTER 11

Searching the Tritryp Genomes for Drug Targets

Peter J. Myler*

Abstract

The recent publication of the complete genome sequences of *Leishmania major*, *Trypanosoma brucei* and *Trypanosoma cruzi* revealed that each genome contains 8300-12,000 protein-coding genes, of which ~6500 are common to all three genomes, and ushers in a new, post-genomic, era for trypanosomatid drug discovery. This vast amount of new information makes possible more comprehensive and accurate target identification using several new computational approaches, including identification of metabolic "choke-points", searching the parasite proteomes for orthologues of known drug targets, and identification of parasite proteins likely to interact with known drugs and drug-like small molecules. In this chapter, we describe several databases (such as GENEDB, BRENDA, KEGG, METACYC, the THERAPEUTIC TARGET DATABASE, and CHEMBANK) and algorithms (including PATHOLOGIC, PATHWAY HUNTER TOOL, AND AUTODOCK) which have been developed to facilitate the bioinformatic analyses underlying these approaches. While target identification is only the first step in the drug development pipeline, these new approaches give rise to renewed optimism for the discovery of new drugs to combat the devastating diseases caused by these parasites.

Traditionally, drug discovery in the trypanosomatids (and other organisms) has proceeded from two different starting points: screening large numbers of existing compounds for activity against whole parasites or more focused screening of compounds for activity against defined molecular targets. Most existing anti-trypanosomatids drugs were developed using the former approach, although the latter has gained much attention in the last twenty years under the rubric of "rational drug design". Until recently, one of the major bottlenecks in anti-trypanosomatid drug development has been our ability to identify good targets, since only a very small percentage of the total number of trypanosomatid genes were known. That has now changed forever, with the recent (July, 2005) publication of the "Tritryp" (*Trypanosoma brucei*, *Trypanosoma cruzi* and *Leishmania major*) genome sequences.[1-4] This vast amount of information now makes possible several new approaches for target identification and ushers in a post-genomic era for trypanosomatid drug discovery.

Tritryp Genome Content

According to the latest data released at GeneDB (http://www.genedb.org), the haploid genomes of *T. brucei*, *T. cruzi* and *L. major* encode 9878, ~12000, and 8373 likely protein-coding genes (and pseudogenes), respectively (see Table 1). The gene densities of the two trypanosome genomes are quite similar (300-400 genes/Mb) and somewhat higher than that of *L. major* (250 gene/Mb). The average coding sequence (CDS) is slightly larger in *Leishmania*, as a result

*Peter J. Myler—Seattle Biomedical Research Institute, 307 Westlake Ave N., Suite 500, Seattle, Washington 98109-5219, USA. Email: peter.myler@sbri.org

Drug Targets in Kinetoplastid Parasites, edited by Hemanta K. Majumder.

Table 1. Tritryp genome statistics

	T. brucei	*T. cruzi*	*L. major*
Protein-coding genes	8599[a]	~10,000[b]	8302
Pseudogenes	1279	~2000[c]	71
Average CDS length (bp)	1,592	1,513	1,901
Average inter-CDS size	1,279	1,024	2,045
Gene density (gene/Mb)	317	385	252
Function known	5%	n.d.[d]	4%
Function inferred	38%	43%	28%
Hypothetical, conserved	51%	48%	56%
Hypothetical, species-specific	6%	9%	8%
Orthologues in all Tritryps	73%	54%	80%
Tb+Tc only	5%	4%	-
Tb+Lm only	1%	-	1%
Tc+Lm only	-	4%	6%
Species-specific	21%	38%	13%

a) Excludes 612 genes annotated as hypothetical protein, unlikely. b) Total number in both haplotypes is 18,980. c) Number in both haplotypes is 3,590. d) Not determined.

of small sequence insertions relative to the trypanosomes, but the lower gene density in *Leishmania* is mostly explained by its larger inter-CDS regions. Each species contains a number of gene families of varying size. Predicted functions have been ascribed to ~40% of the protein-coding genes, but this has been confirmed experimentally for only ~5% of the proteins. Most of the remaining genes encode conserved hypothetical proteins, of which slightly more than half are found only in trypanosomatids. Interestingly, ~2-3% of the Tritryp proteins are related to those found in prokaryotes but not other eukaryotes. At least some of these appear to have arisen from horizontal gene transfer, and may represent excellent candidates for drug targets. The Tritryp genomes display a remarkable degree of synteny, with ~75% of the genes in *L. major* having orthologues in both other species and >90% of these occurring in the same genomic context (see Table 1). The proteins within this Tritryp "core" proteome exhibit an average 57% identity between *T. brucei* and *T. cruzi*, and 44% identity between *L. major* and the two other trypanosomes, reflecting the expected phylogenetic relationships.[5,6] Interestingly, substantially fewer orthologues are shared only between *L. major* and *T. brucei* than between *L. major* and *T. cruzi*, perhaps reflecting the common intracellular environment of their mammalian stages.

However, all three genomes contain a significant number of species-specific genes, which account for ~21% and 38% of the protein-coding genes in *T. brucei* and *T. cruzi*, respectively, but only ~13% of the *L. major* genes. These species-specific genes (and pseudogenes) mostly encode large families of surface proteins, exemplified by the variant surface glycoproteins (VSGs) and Procyclic Acidic Repetitive Proteins (EP/PARP/procyclin) of *T. brucei*; the trans-sialidases, dispersed gene family protein 1 (DGF-1), mucins, and mucin-associated surface proteins (MASPs) of *T. cruzi*; and the amastins and promastigote surface antigens (PSA-2) of *L. major*. In addition to these species-specific genes, all three species demonstrate differential paralogous gene expansion or contraction, with the ESAG4 adenylate/guanylate cyclases and leucine-rich repeat proteins being over-represented in *T. brucei*; GP63 surface proteases and recombination hot spot (RHS) proteins in *T. cruzi*; and mitochondrial carrier protein, ATP-Binding Cassette (ABC) transporters, and Heat Shock Protein (HSP) 90 gene families in *L. major*. Many of these species-specific genes or paralogous expansions occur in telomeric and sub-telomeric gene clusters, possibly reflecting similar strategies used for immune evasion.

Transcription and RNA processing in the trypanosomatids is quite different from that in other eukaryotes,[7] with unique or unusual processes such as large polycistronic gene clusters,[8-10] RNA polymerase I-mediated transcription of some protein-coding genes,[11,12] and trans-splicing.[13] While annotation of the Tritryp genomes uncovered most of the expected RNAP polymerase subunits, there was a dearth of transcription factors normally involved in regulation of transcription initiation by other eukaryotes.[3] However, recent experiments have identified several highly divergent transcription factors in *T. brucei*,[14-17] suggesting that Tritryp transcription initiation may represent an ancestral, less sequence-specific, mechanism mostly replaced in other eukaryotes by the archetypal TATA-containing promoters. Conversely, the paucity of Tritryp genes encoding transcriptional regulators is offset by an abundance of proteins with RNA binding motifs,[18] consistent with their reliance on post-transcriptional models of gene regulation.[19]

DNA replication in trypanosomatids also appears to differ significantly from that in higher eukaryotes, with only one of the six subunits typically found in the eukaryotic replication origin complex being identified.[2] There are also substantial differences in the mitochondrial replication machinery, since the complexity of the kinetoplast DNA (the trypanosomatid equivalent of a mitochondrial genome) structure dictates an unusual replication mechanism.[20]

Bioinformatic analyses of the Tritryp genomes suggests that they lack several classes of signaling molecules found in other eukaryotes, including serpentine receptors, heterotrimeric G proteins, most classes of catalytic receptors, SH2 and SH3 interaction domains, and regulatory transcription factors, but that they do possess a large and complex set of protein kinases and protein phosphatases.[2,21] However, the distribution of protein kinase classes differs from that in other organisms; with no tyrosine kinases (other than dual specificity kinases), receptor kinases or TKL and RGC group kinases. Since the trypanosomatids have complicated life cycles in different hosts, it is likely that these kinases play important roles in regulating their response to changes in these different environments.

Computational Approaches for Drug Target Selection

The experience gained by the pharmaceutical industry during the last few decades of drug development has lead to the postulation of a number of selection criteria for successful drug target identification.[22] In the context of the trypanosomatids, these criteria include selectivity (i.e., the parasite target is absent from, or substantially different in, the host); "druggability" (the target structure has a small molecule-binding pocket); suitable biochemical properties (the target has a low turnover rate and/or catalyzes a rate-limiting step within a pathway); validation (the target is essential for growth and/or survival in the mammalian stage of the parasite lifecycle); "assayability" (specific, inexpensive and high-throughput screens are available using in vitro expressed target); and low potential for development of drug resistance (absence of different isoforms or alleles and/or biochemical "bypass" reactions). With these criteria in mind, several bioinformatic approaches have been proposed, which take advantage of the availability of the complete genome sequences described above to accelerate progress in developing effective clinical interventions for the important diseases caused by these parasites.

Analysis of the Tritryp genomes has provided a comprehensive view of the parasites' metabolic potential by identifying numerous common and species-specific metabolic and transport processes. Manual examination of metabolic maps identified a number of pathways that appear to be especially amenable to potential chemotherapeutic intervention; including glycolysis, the electron transport chain, the urea cycle, the glyoxylase pathway and associated trypanothione metabolism, glycosylphosphatidylinositol (GPI) anchor biosynthesis, fatty acid biosynthesis, as well as the ergosterol and isoprenoid biosynthetic pathway.[1] Since the particulars of target identification and drug development for each of these pathways (and others) have been described in detail in several of the accompanying chapters and elsewhere,[23-26] they will not be further explored here. Instead, several different computational attempts to catalogue metabolic pathways and identify "choke-points" will be described.

Databases of Tritryp Metabolism

BRENDA (BRaunschweig ENzyme DAtabase) is a comprehensive collection of enzyme and metabolic information (http://www.brenda.uni-koeln.de), including Enzyme Commission (EC) classification and nomenclature, reaction and specificity, function and structure, isolation and stability, as well as links to primary literature references. The database is now based on a controlled vocabulary and ontology for some information fields, and search tools include EC and taxonomy-tree browsers, a chemical substructure search engine for ligand structure, and a thesaurus for ligand names. BRENDA contains more than 100,000 enzymes representing 4060 different EC numbers from about ~10000 different organisms. There are currently (as of September, 2006) 842 entries for *T. brucei*, 751 for *T. cruzi* and 607 for *L. major*.

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a suite of databases and associated software, designed to integrate current knowledge of genes and proteins (GENES database), chemical compounds and reactions (LIGAND), metabolic, regulatory and interaction networks (PATHWAY), and ontologies (BRITE). Biological systems are represented in KEGG by nested graphs, which are used for pathway reconstruction and functional inference, and line graphs, which form the basis for integrating genome and chemical information with the networks. The BRITE database provides the pathway reconstruction through a series of functional hierarchies and represents the logical foundation for the KEGG project. KEGG maintains a gene catalogue of sequenced genomes and maps them onto 301 manually drawn and curated reference pathways.[27-31] Currently, there are 83, 90, and 89 entries in the PATHWAY database for *T. brucei*, *T. cruzi* and *L. major*, respectively, mostly describing metabolic pathways.

The BIOCYC collection of Pathway/Genome Databases (PGDBs) provides electronic reference sources on the pathways and genomes of more than 200 different organisms (http://biocyc.org). The databases within the BIOCYC collection are organized into tiers according to the amount of manual review and updating they have received. Tier 1 PGDBs are created through intensive manual efforts, and receive continuous updating. EcoCyc, which describes *Escherichia coli* K-12, is the only organism-specific Tier 1 database. Tier 2 PGDBs are computationally generated using PATHOLOGIC software,[32,33] and have undergone moderate amounts of review and updating. There are currently 12 databases in Tier 2, including HUMANCYC and PLASMOCYC (which describes the malaria parasite, *Plasmodium falciparum*). Tier 3 databases are computationally generated by the PATHOLOGIC program, and have undergone no review and updating.[34] There are 191 PGDBs in Tier 3, representing mostly bacterial genomes. The individual BIOCYC web-sites can be used to visualize single or multiple metabolic pathways, including a complete metabolic map of the organism. An OMICS VIEWER can be used to analyze gene expression, proteomics, or metabolomics data to produce animated views of time-course gene-expression experiments. There are currently no BIOCYC PGDBs for any of the trypanosomatid genomes, although it should be relatively straightforward to generate Tier 3 databases using the PATHOLOGIC software.[32] Other programs are also available for genome-scale reconstruction of metabolic networks.[35-38] However, since this process is largely dependent on sequence-based homology searches to identify the enzymes and the Tritryp genomes are quite divergent from other eukaryotes, considerable manual curation will probably be necessary to obtain truly accurate representations of the metabolic networks in these organisms.

While most of the individual PGDBs within BIOCYC represent species-specific databases, METACYC (http://metacyc.org) is a collection of metabolic pathways and enzymes from more than 240 organisms (mostly bacteria and plants). The goal of METACYC is to represent every experimentally elucidated metabolic pathway, reaction, and chemical compound, as well as the genes encoding the enzymes that catalyze the reactions involved.[39] As well as being used as a reference source to look up individual facts, METACYC facilitates computational studies of the metabolism, such as design of novel biochemical pathways for biotechnology, studies of evolution of metabolic pathways, and simulation of metabolic pathways. Additionally, desktop software is available for comparing the overall metabolic maps, specific pathways and genomic maps of two organisms.

Identification of Metabolic "Choke-Points"

Careful manual examination of a metabolic pathway can identify metabolic "choke-points", i.e., the enzyme(s) which is (are) uniquely necessary to produce a critical metabolite. Obviously, choke-points in pathways that result in metabolites critical for parasite survival would make excellent potential targets for development of novel anti-trypanosomatid drugs. The PATHWAY HUNTER TOOL (http://www.pht.uni-koeln.de) uses an extended form of graph theory (in which enzymes are represented by edges between nodes representing metabolites) to identify choke-points and rank them according to their "load".[40,41] Load is defined as the ratio of the number of shortest paths through the enzyme and nearest neighbors attached to it, compared to the average values for these properties in the entire network. Comparison of pathogen (trypanosomatid) and host (human) metabolic networks could be used to identify highly ranked choke-points that are unique to the parasite or are ranked much lower in the host.

Another computational approach for identification of metabolic enzymes as drug targets involves the concept of minimal cut sets, which are defined as the minimal set of reaction in a network whose inactivation will definitively lead to a failure in a particular network function.[42] Screening parasite metabolic networks for all possible minimal cut sets and identification of those which are small (i.e., contain few enzymes) and not present in the host could serve to identify potential drug targets.

The approaches outlined above are designed to identify targets that meet only some of the criteria outlined at the beginning of this section; namely they have suitable biochemical properties, are likely to be essential for the parasite, and are sufficiently different from any host homologue. However, alternative approaches seek to make use of the finding that successful drugs have specific structural and physicochemical properties that allow them to be efficacious, bioavailable, and safe. These properties are exemplified by Lipinski's so-called "rule of five".[43] This has lead to the concept of "druggable" proteins, based on their ability to bind potentially effective drug-like small molecules.[44-46] Thus, it makes sense to search the Tritryps genomes for proteins that are likely to meet these criteria. Two different approaches have been proposed for developing computational solutions to this problem: searching the genome for proteins with similar properties to known drug targets in other organisms (primarily humans) and direct interrogation of the parasite proteins for their likelihood to bind drug-like chemicals.

Searching for Parasite Orthologues of Known Drug Targets

The Therapeutic Target Database (TTD) (http://xin.cz3.nus.edu.sg/group/cjttd/ttd.asp) represents a comprehensive and publicly available attempt to catalogue information about all the currently known protein and nucleic acid targets described in the literature.[46,47] The database also contains information about the drugs and ligands directed at these targets, as well as corresponding disease conditions. This database currently contains 1535 targets and 2107 drugs/ligands, including 19 entries listing potential anti-trypanosomatid use. The most simplistic approach for searching the Tritryp genomes for potential targets similar to these existing targets would be to carry out BLASTP or PSIBLAST searches of the Tritryp protein databases to identify parasite proteins with significant sequence similarity to those in the TTD. The resulting list of parasite proteins would need to be subsequently winnowed down by removing those that are too similar to the human orthologues and/or are similar to proteins involved in more than two pathways in humans, since drugs against these are likely to have deleterious effects on the human host. However, given what we know about the imprecise nature of the relationship between protein sequence and structure, it is likely that this method will have a significant false negative rate (i.e., it will miss many potentially useful targets because they won't have sufficient sequence similarity). Statistical learning methods, such as support vector machines (SVM) and neural networks, have recently enjoyed considerable success for prediction of protein structure and may be useful for identifying targets missed by simple BLAST searching. A SVM method has been used to screen the human and HIV genome for druggable proteins, with a promising degree of accuracy.[46,48,49] Similar methods could be used to screen the Tritryp genomes.

Matching Drug-Like Chemicals to Parasite Proteins

Algorithms such as AUTODOCK[50] have been used for some time to predict small molecules that will potentially fill protein ligand-binding pockets, as a first step in rational drug design. This process has been reversed to some extent by using docking software with integrated molecular dynamics simulation to predict which drugs are likely to bind (and inhibit) proteases from human coronavirus,[51] cytomegalovirus,[52] and human immunodeficiency virus (HIV).[53] A recent publication describes the use of this method to screen 2500 compounds in the CHEMBANK database (http://chembank.broad.harvard.edu) against 13 proteins from *Plasmodium falciparum* whose structure had been determined by X-ray crystallography.[54] This approach found that the *Ki*s predicted for three existing anti-malarial drugs compared well with their known values and that their predicted inhibitory activity ranked in the top 5th percentile of all tested drugs. Another 20 drugs were predicted to have multi-target activity, i.e., they showed high affinity with two or more proteins. Multi-target drugs are attractive because they are less likely to encounter problems with development of drug resistance. It should be possible to screen the Tritryp proteome for multi-target drugs using a similar approach. Obviously, one major constraint is the availability (or lack thereof) of trypanosomatid proteins with known structure. Currently, the protein structure database (PDB) contains 79 nonredundant structures from the genera *Leishmania* or *Trypanosoma*. However, this number has been increasing rapidly over the last few years due to the efforts of the Structural Genomics of Pathogenic Protozoa (SGPP) consortium (http://www.sgpp.org) and is likely to increase further in the near future.

Conclusion

The recent completion of the Tritryp genome sequencing project provides an unprecedented opportunity for development of novel anti-trypanosomatid chemotherapeutic agents. The identification of more than 8000 new protein-coding genes, many of which are shared between the *Leishmania* and *Trypanosoma* genera, vastly expands the potential drug targets available for investigation. In fact, the situation has gone from a relative dearth of useful targets to an embarrassment of riches, with far more potential targets available than can possibly be studied in detail. In this chapter, we have described several different computational approaches that should be useful in reducing this smorgasbord of genes to a manageable number of high-value targets, which will form the basis of detailed biological and pharmacological investigation. Of course, target identification is only the first stages in the lengthy and expensive process of drug development; with steps such as target validation, lead identification and optimization, as well as preclinical pharmacological screening, necessary before a potential drug can enter clinical trials. Nevertheless, these bioinformatic methods hold great promise in being able to identify targets (and potential lead compounds in some cases) which have a higher probability of successful drug development than traditional methods. While only time will reveal the validity of this promise, we hope that this advent of the post-genomics era for trypanosomatid biology heralds a renaissance in the discovery of much needed new drugs for the devastating diseases caused by these parasites.

References

1. Berriman M, Ghedin E, Hertz-Fowler C et al. The genome of the African trypanosome, Trypanosoma brucei. Science 2005; 309:416-422.
2. El-Sayed NMA, Myler PJ, Bartholomeu D et al. The genome sequence of Trypanosoma cruzi, etiological agent of Chagas' disease. Science 2005; 309(5733):409-415.
3. Ivens AC, Peacock CS, Worthey EA et al. The genome of the kinetoplastid parasite, Leishmania major. Science 2005; 309(5733):436-442.
4. El-Sayed NMA, Myler PJ, Blandin G et al. Comparative genomics of trypanosomatid parasitic protozoa. Science 2005; 309(5733):404-409.
5. Haag J, O'hUigin C, Overath P. The molecular phylogeny of trypanosomes: Evidence for an early divergence of the Salivaria. Mol Biochem Parasitol 1998; 91(1):37-49.
6. Stevens JR, Noyes HA, Schofield CJ et al. The molecular evolution of Trypanosomatidae. Adv Parasitol 2001; 48:1-56.

7. Campbell DA, Thomas S, Sturm N. Transcription in kinetoplastid protozoa: Why be normal? Microbes Infect 2003; 5(13):1231-1240.
8. Myler PJ, Audleman L, deVos T et al. Leishmania major Friedlin chromosome 1 has an unusual distribution of protein-coding genes. Proc Natl Acad Sci USA 1999; 96(6):2902-2906.
9. Martinez-Calvillo S, Yan S, Nguyen D et al. Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions within a single region. Mol Cell 2003; 11(5):1291-1299.
10. Martinez-Calvillo S, Nguyen D, Stuart K et al. Transcription initiation and termination on Leishmania major chromosome 3. Eukaryot Cell 2004; 3(2):506-517.
11. Vanhamme L, Pays E. Control of gene expression in trypanosomes. Microbiol Rev 1995; 59(2):223-240.
12. Lodes MJ, Merlin G, deVos T et al. Increased expression of LD1 genes transcribed by RNA polymerase I in Leishmania donovani as a result of duplication into the rRNA gene locus. Mol Cell Biol 1995; 15(12):6845-6853.
13. Perry K, Agabian N. mRNA processing in the Trypanosomatidae. Experientia 1991; 47:118-128.
14. Das A, Zhang Q, Palenchar JB et al. Trypanosomal TBP functions with the multisubunit transcription factor tSNAP to direct spliced-leader RNA gene expression. Mol Cell Biol 2005; 25(16):7314-7322.
15. Schimanski B, Nguyen TN, Günzl A. Characterization of a multisubunit transcription factor complex essential for spliced-leader RNA gene transcription in Trypanosoma brucei. Mol Cell Biol 2005; 25(16):7303-7313.
16. Palenchar JB, Liu W, Palenchar PM et al. A divergent transcription factor TFIIB in trypanosomes is required for RNA polymerase II-dependent SL RNA transcription and cell viability. Eukaryot Cell 2006; 5(2):293-300.
17. Schimanski B, Brandenburg J, Nguyen TN et al. A TFIIB-like protein is indispensable for spliced leader RNA gene transcription in Trypanosoma brucei. Nucl Acids Res 2006; 34(6):1676-1684.
18. Anantharaman V, Aravind L, Koonin EV. Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. Curr Opin Chem Biol 2003; 7(1):12-20.
19. Clayton CE. Life without transcriptional control? From fly to man and back again. EMBO J 2002; 21(8):1881-1888.
20. Klingbeil MM, Motyka SA, Englund PT. Multiple mitochondrial DNA polymerases in Trypanosoma brucei. Mol Cell 2002; 10(1):175-186.
21. Parsons M, Worthey EA, Ward PN et al. Comparative analysis of the kinomes of three pathogenic trypanosomatids: Leishmania major, Trypanosoma brucei and Trypanosoma cruzi. BMC Genomics 2005; 6(1):127.
22. Pink R, Hudson A, Mouries MA et al. Opportunities and challenges in antiparasitic drug discovery. Nat Rev Drug Discov 2005; 4(9):727-740.
23. Fairlamb AH. Chemotherapy of human African trypanosomiasis: Current and future prospects. Trends Parasitol 2003; 19(11):488-494.
24. Lee SH, Stephens JL, Paul KS et al. Fatty Acid synthesis by elongases in trypanosomes. Cell 2006; 126(4):691-699.
25. Albert MA, Haanstra JR, Hannaert V et al. Experimental and in silico analyses of glycolytic flux control in bloodstream form Trypanosoma brucei. J Biol Chem 2005; 280(31):28306-28315.
26. Lakhdar-Ghazal F, Blonski C, Willson M et al. Glycolysis and proteases as targets for the design of new anti-trypanosome drugs. Curr Top Med Chem 2002; 2(5):439-456.
27. Goto S, Nishioka T, Kanehisa M. LIGAND: Chemical database for enzyme reactions. Bioinformatics 1998; 14(7):591-599.
28. Kanehisa M. A database for post-genome analysis. Trends Genet 1997; 13(9):375-376.
29. Kanehisa M, Goto S, Hattori M et al. From genomics to chemical genomics: New developments in KEGG. Nucleic Acids Res 2006; 34(Database issue):D354-D357.
30. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000; 28(1):27-30.
31. Kanehisa M, Goto S, Kawashima S et al. The KEGG resource for deciphering the genome. Nucleic Acids Res 2004; 32(Database issue):D277-D280.
32. Karp PD, Paley S, Romero P. The pathway tools software. Bioinformatics 2002; 18(Suppl 1):S225-S232.
33. Yeh I, Hanekamp T, Tsoka S et al. Computational analysis of Plasmodium falciparum metabolism: Organizing genomic information to facilitate drug discovery. Genome Res 2004; 14(5):917-924.
34. Karp PD, Ouzounis CA, Moore-Kochlacs C et al. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Res 2005; 33(19):6083-6089.
35. Ma H, Zeng AP. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics 2003; 19(2):270-277.

36. Covert MW, Schilling CH, Famili I et al. Metabolic modeling of microbial strains in silico. Trends Biochem Sci 2001; 26(3):179-186.
37. Gaasterland T, Selkov E. Reconstruction of metabolic networks using incomplete information. Proc Int Conf Intell Syst Mol Biol 1995; 3:127-135.
38. Overbeek R, Larsen N, Pusch GD et al. WIT: Integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res 2000; 28(1):123-125.
39. Krieger CJ, Zhang P, Mueller LA et al. MetaCyc: A multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 2004; 32(Database issue):D438-D442.
40. Rahman SA, Schomburg D. Observing local and global properties of metabolic pathways: 'Load points' and 'choke points' in the metabolic networks. Bioinformatics 2006; 22(14):1767-1774.
41. Rahman SA, Advani P, Schunk R et al. Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). Bioinformatics 2005; 21(7):1189-1193.
42. Klamt S, Gilles ED. Minimal cut sets in biochemical reaction networks. Bioinformatics 2004; 20(2):226-234.
43. Lipinski CA, Lombardo F, Dominy BW et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 2001; 46(1-3):3-26.
44. Hopkins AL, Groom CR. The druggable genome. Nat Rev Drug Discov 2002; 1(9):727-730.
45. Hardy LW, Peet NP. The multiple orthogonal tools approach to define molecular causation in the validation of druggable targets. Drug Discov Today 2004; 9(3):117-126.
46. Zheng CJ, Han LY, Yap CW et al. Therapeutic targets: Progress of their exploration and investigation of their characteristics. Pharmacol Rev 2006; 58(2):259-279.
47. Chen X, Ji ZL, Chen YZ. TTD: Therapeutic Target Database. Nucleic Acids Res 2002; 30(1):412-415.
48. Han L, Cui J, Lin H et al. Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity. Proteomics 2006; 6(14):4023-4037.
49. Cai CZ, Han LY, Ji ZL et al. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res 2003; 31(13):3692-3697.
50. Goodsell DS, Morris GM, Olson AJ. Automated docking of flexible ligands: Applications of AutoDock. J Mol Recognit 1996; 9(1):1-5.
51. Jenwitheesuk E, Samudrala R. Identifying inhibitors of the SARS coronavirus proteinase. Bioorg Med Chem Lett 2003; 13(22):3989-3992.
52. Jenwitheesuk E, Samudrala R. Virtual screening of HIV-1 protease inhibitors against human cytomegalovirus protease using docking and molecular dynamics. AIDS 2005; 19(5):529-531.
53. Jenwitheesuk E, Wang K, Mittler JE et al. PIRSpred: A web server for reliable HIV-1 protein-inhibitor resistance/susceptibility prediction. Trends Microbiol 2005; 13(4):150-151.
54. Jenwitheesuk E, Samudrala R. Identification of potential multitarget antimalarial drugs. JAMA 2005; 294(12):1490-1491.