Journal of
Proteomics & Bioinformatics

Chen J Proteomics Bioinform 2011, 4:4
DOI: 10.4172/jpb.1000172

Research Article Open Access

Reconstruction of Local Biochemical Reaction Network Based on Human Chromosome 9 Sequence Data

Biqing Chen*

College of Life Sciences, Zhejiang Unviersity, China

Abstract

Metabolic networks are complex and highly interconnected, thus systems-level computational approaches are required to elucidate and to understand metabolic genotype–phenotype relationships. This paper has manually reconstructed the local human metabolic network based on DNA sequence data of human chromosome nine. Herein the paper describes the reconstruction process and discusses how the resulting chromosome-scale (or local) network differs from genome-scale ones. The underestimated results have revealed many gaps in the current understanding of human metabolism that require future experimental investigation. They also suggest possible problems arising from local reconstruction based on partial genome data. The study suggests further applications enabled by reconstruction of human metabolic network. The establishment of this network represents a step toward genome-scale human systems biology.

Keywords: Metabolism; Prediction; Reconstruction; Partial genome

Introduction

After the Human Genome Project was finished at the beginning of 21st century, it seems that we have our own destiny under control. However, what is the next step to deal with such a large pool of data so that they are meaningful to people? The direct approach to understanding the complex processes encoded by the human genome is studying gene products' function, assigning these enzyme products to biochemical pathways and reconstructing biochemical networks. These biochemical pathways define regulated sequences of biochemical transformations. It is a first step towards quantitative modelling of metabolism. An individual's metabolism is determined by one's genetics, environment, and nutrition. Hopefully, with the available human genome sequence and its annotation [1-3], the human body's metabolic network can be reconstructed. Numerous metabolic genes and enzymes have been individually studied for decades; however, these results are dispersed without integrated understanding. The procedure for integrating these diverse data types to form a network reconstruction and predictive model is well established for microorganisms [4] and has been applied to mouse hybridomas [5]. Such *in silico* models have enabled hypothesis-driven biology, including the prediction of the outcome of adaptive evolution [6-10] and the identification and discovery of candidates for missing metabolic functions that were subsequently experimentally verified [11]. Because metabolic networks are more complex in mammals than in single-celled organisms, there is likely to be an even greater opportunity for the use of computational models to understand the basis of normal and abnormal cellular function [12]. At the same time, reconstructing biochemical reaction networks in mammals is also more complex and tougher. Assignment of genes to pathways also permits a validation of the human genome annotation because patterns of pathway assignments spotlight likely false-positive and false-negative genome annotations.

Previous researches have reconstructed global human metabolic network based on genomic data. *Homo sapiens* Recon 1 is a comprehensive literature-based genome-scale metabolic reconstruction that accounts for the functions of 1,496 ORFs, 2,004 proteins, 2,766 metabolites, and 3,311 metabolic and transport reactions [12]. Another computational prediction of human metabolic pathways from the complete human genome assigns 2,709 human enzymes to 896 bioreactions and 622 of the enzymes are assigned roles in 135 predicted metabolic pathways [13].

This paper presents the reconstruction of the local human metabolic network only with information from human chromosome 9(referred as chr9 in this paper). Bottom-up reconstruction method was used in this paper. The comprehensive database-based chromosome-scale metabolic reconstruction, named as Rec 9, accounts for 53 ORFs, 53 proteins, 4 nonenzymatic proteins, 16 metabolic enzymes(7 of them are redundant), and 9 natural metabolic and transport reactions. Rec 9 (*i*) enhances our understanding of gene inter-locking rules and the relationship between inter-locked genes' products, (*ii*) facilitates the computational interrogation of the overall properties of the human metabolic network, and (*iii*) provides supplemental context for analysis of "-omics" data sets.

Methods

The complete DNA sequence of human chr9 was downloaded from NCBI GeneBank, in FASTA format. Prediction of genes on chr9 and their corresponding peptides was conducted by GENSCAN. Considering the maximum DNA length limitation, the whole sequence of chr9 was randomly divided into two parts, each part was processed by GENSCAN separately. Set HumanIso.Smat as the parameter matrix. SAPS was utilized to elucidate proteins' physiochemical properties. Secondary structure, together with functional domain, was analyzed using several online tools, namely 9aaTAD, Scratch Protein Predictor, NetSurfP, SOPMA, PeptideCutter, and ELM. Subcellular location of each protein was also analyzed with TargetP.

Because these predicted proteins are uncharacterized and unknown, the reactions they are involved in can't be searched directly

***Corresponding author:** Biqing Chen, College of Life Sciences, Zhejiang University, Tel: 8613958101344; E-mail: bq_chen@qq.com

Received February 23, 2011; **Accepted** April 11, 2011; **Published** April 15, 2011

Citation: Chen B (2011) Reconstruction of Local Biochemical Reaction Network Based on Human Chromosome 9 Sequence Data. J Proteomics Bioinform 4: 087-090. doi:10.4172/jpb.1000172

in databases. To solve this problem, two ways were suggested. One is de novo discover, that is using some analyzing tools to predict potential functions from protein sequence information. In the other method, these sequences were blasted and their functions were regarded the same as their identities' (similar known proteins') functions. The blast process used BLASTP 2.2.23, chose UniProtKB (Protein) as the database and blosum62 as the scoring matrix. All the candidates selected under certain criteria from the blast results were searched against UniProtKB for detailed description. Each candidate was manually examined to see whether it has certain characters. Candidates with similar function were regarded as one protein. A second selection was done then based on the information richness and redundancy of these candidate proteins.

Reactions these predicted proteins involved in were defined as those catalyzed by the enzymes determined in the final selection. Thus, structural proteins and proteins with unknown function were not examined in the reaction prediction step. Reactions involving predicted proteins as substrates were not included since it is rare for a gene product to be a metabolite. Reactions were obtained from KEGG ENZYME database and BRENDA.

KEGG LIGAND was used to map all the reactions predicted before, in detail, Pathway Mapper was employed. Besides, each reaction was also searched against KEGG PATHWAY database to view the whole map of pathway it involved in. All the predicted proteins were also searched against EMBL REACTOME database to directly find out whether they are involved in any known pathways.

All the processes were limited to *Homo Sapiens* if species specification was possible.

Results

The whole DNA sequence of chr9 is 1246910 bp(divided into 939400bp, referred as s1; 307510bp, referred as s2). The average G+C is 41.04%(42.11% for s1; 39.96% for s2). GENSCAN predicted 53 genes (Supplementary Information 1), including information about their structures. GENSCAN results also predicted 53 peptides accordingly (Supplementary Invitation 2), so there is no need to use another tool, such as Transeq, to translate nucleotides into proteins.

2 predicted proteins were located in mitochondria; 5 were associated with secretary pathway (Supplementary Information 3). Physiochemical property analysis showed these predicted peptides vary in amino acids composition, thus differ in charge distribution (Supplementary Information 4). Nevertheless, it is also necessary to infer secondary structures of these predicted proteins for function is usually associated with protein' higher structures. The results by different online tools were compared and final results of secondary structure (Supplementary Information 5) came from Scratch Protein Predictor (ExPASy).

The blast result is a list of proteins in the order of increasing E-value. The selection of similar proteins to represent the predicted protein is a problem. To which degree of similarity can we define the function of a predicted protein the same as that of its identity? Considering some relative numbers in alignment, candidates were selected with identity higher than 80% and relatively low E-value in this study. 166 protein candidates, corresponding to 24 predicted peptides, were selected under this criterion (supplementary data not shown here). 8 of these predicted proteins had matches with 100% identity, and E-values were also low enough so that we can equate these predicted proteins to their identical matches. However, most of these 166 candidates were uncharacterized proteins or isoforms of each other. After filtering according to their ontology and information from UniProtKB, 13 distinctive proteins were determined (Table 1), of which 9 were enzymes, which could be subdivided into 7 metabolic enzymes, plus 2 nonmetabolic enzymes (including enzymes whose substrates are macromolecules, such as protein kinases and DNA polymerases). The 'Unmatched' row includes predicted proteins with no character information, while the remaining 4 nonenzymatic proteins are listed in the 'Nonenzyme' row.

Some proteins' descriptions were not well-characterized that one protein may retrieve several enzymes with different EC number but similar function. In such cases, extra alignments between the predicted protein and its identities were done. The enzyme with the highest score was chosen. Sometimes several predicted proteins were assigned to one enzyme, which means several genes have the same function. 30.2% of the predicted genes coded enzymes, correspond to 28.75% of the whole chr9 DNA.

KEGG assigned one reaction to one enzyme, while BRENDA offered as many reactions as possible from literature study. Thus, sometimes one gene product was matched to more than one reaction, as happens with multifunctional enzymes. All together 82 reactions (52 of them had missing information on products) were predicted besides

Type	Pro#	Identity's Name	Main Reaction or Function
Unmatched	5,8,27,46		
Nonenzyme	2	transmembrane protein	integral to membrane
	13	FAM122A	
	39	FAM189A2	
	48	forkhead box protein D4	transcription factor
Nonmetabolic Enzyme	22	cAMP-dependent protein kinase γ subunit	ATP + a protein = ADP + a phosphoprotein
	33,41,44,45	deoxynucleoside-triphosphate:DNA deoxynucleotidyltransferase (RNA-directed)	deoxynucleoside triphosphate+DNAn = diphosphate+DNAn+1
Metabolic Enzyme	7	chondroitin-D-glucuronate 5-epimerase	chondroitin D-glucuronate = dermatan L-iduronate
	18	ATP:1-phosphatidyl-1D-myo-inositol-4-phosphate 5-phosphotransferase	ATP+1-phosphatidyl-1D-myo-inositol-4-phosphate=ADP+1-phosphatidyl-1D-myo-inositol-4,5-Bisphosphate
	24,25,26	Frataxin	$4Fe^{2+}+4H^{+}+O_2 = 4Fe^{3+}+2H_2O$
	28	GTP:AMP phosphotransferase	GTP+AMP = GDP+ADP
	31	tight junction protein 2	tight junction; ATP+GMP=ADP+GDP
	43,47	RTP:adenosylcobinamide phosphotransferase	RTP+adenosylcobinamide=adenosylcobinamide phosphate+RDP [RTP is either ATP or GTP]
	51,52	alpha-D-glucose 1,6-phosphomutase	alpha-D-glucose 1-phosphate = D-glucose 6-phosphate

Pro#, the ID number of predicted protein, in this study we set the order of the predicted proteins resulting from GENSCAN as 1~53.

Table 1: Identified predicted proteins and corresponding function from BLAST results.

Pro#	Pathways Involved	Reactions Related
22	glucose metabolism	PFKFB1 dimer+ATP=phosphor PFKFB1 dimer+ADP
	lipid digestion,mobilization, and transport	perilipin+ATP=phosphorylated perilipin+ADP; hormone sensitive lipase+ATP=phosphorylated hormone sensitive lipase+ADP
	hemostasis	
	opioid signalling	PKA tetramer + 3',5'-cyclic AMP = PKA catalytic subunit + cAMP : PKA regulatory subunit; ATP+CREB=ADP+phosphor-CREB
	NGF signalling	ATP+CREB=ADP+phosphor-CREB
	glucagon signaling	ChREBP protein+ATP= pChREBP(Thr 666)+ADP; pChREBP(Thr 666)+ATP= pChREBP(Ser 196, Thr 666)+ADP; PFKFB1 dimer+ATP=phosphor PFKFB1 dimer+ADP
	Hormone-sensitive lipase (HSL)-mediated triacylglycerol hydrolysis	
	regulation of insulin secretion	
	Rap1 signalling	Rap1 GTPase-activating protein 2+ATP=p(S7)-Rap1 GTPase- activating protein 2+ADP
	regulation of water balance by renal aquaporins	aquaporin-2 tetramer+ATP=phosphorylated aquaporin-2 tetramer+ADP
28	megakaryocyte development and platelet production(hemostasis)	GTP+AMP = GDP+ADP
31	Apoptosis	ZO-2→proteolytically cleaved ZO-2(acted by caspase)

Pro#, the ID number of predicted protein, in this study we set the order of the predicted proteins resulting from GENSCAN as 1~53.

Table 2: REACTOME results of pathways in which predicted proteins involved.

Pro#	Pathways involved
18	Metabolic pathways; Phosphatidylinositol signaling system; Inositol phosphate metabolism;
31	Metabolic pathways; Purine metabolism
43,47	Metabolic pathways; Porphyrin and chlorophyll metabolism
24,25,26	Porphyrin and chlorophyll metabolism

Pro#, the ID number of predicted protein, in this study we set the order of the predicted proteins resulting from GENSCAN as 1~53.

Table 3: KEGG Pathway results of pathways in which predicted proteins involved.

the main 9 ones. Most of them were reactions recorded in BRENDA in the enzymatic reaction list of frataxin(=ferroxidase) and cAMP-dependent protein kinase γ subunit(PKAγ).

REACTOME found 3 predicted proteins involved in 11 known pathways (Table 2). Among them, one protein was multifunctional, the other two were specific to one pathway each. KEGG Pathway Mapper resulted in 7 pathway maps (Table 3), but the streptomycin biosynthesis and biosynthesis of secondary metabolites pathways obviously do not occur in human beings, thus, they were ignored.

Discussion

It is common to find several genes referred to one enzyme, or at least their products have the same function. One explanation is that these gene products may form a protein complex; therefore, their protein annotations are similar. Another explanation is that these genes code isoenzymes, that is enzymes with similar function but various sequences. This may result from gene duplication at some time. Gene duplications commonly happen in the evolution history, especially on the same chromosome. It is a main type of gene mutation and chromosome variation. According to duplication mechanism, it is quite possible to observe 'one enzyme, several genes' as these predicted genes locate on the same chromosome. The result suggests functional redundancy of interlocked genes may be a common phenomenon in higher organisms, which coordinates with the concept of quantitative trait.

Most predicted reactions involved phosphorylation and dephosphorylation of ATP or GTP, indicating their possible correlations in energy metabolism. However, there's still a lot of missing information, and these reactions are relatively not well-interconnected that a proper network can't be built.

The predicted enzymes encoded by human chromosome 9 and their corresponding reactions are relatively less in quantity, compared to other genome-wide predictions with DNA sizes in consideration. This suggests that prediction based on partial genome data is often problematic. Computational prediction of organism metabolic networks should better use whole genome data. The reason is not clear though. It is possible that this study used genes and proteins predicted simply from DNA sequences as input, while genome-wide network reconstruction researches used information from genome annotation as input. Gene annotations may be better as they are curated and confirmed by literature. Another potential inference from this underestimation is that adding data results in nonlinear increase of information, that is to say, the information encoded by whole genome isn't equal to simple adding of information from partial genome data.

The formulation of an *in silico* model from the reconstruction and initial analysis of the network structure will likely be critical in elucidating underlying mechanisms of disease and identifying treatment strategies by developing cell-, tissue-, and context-specific models and building additional layers of complexity into the framework. Genome-scale microbial metabolic reconstructions have been widely used to successfully perform systems analysis to the point that models resulting from these reconstructions have become tools for hypothesis driven biological discovery [4]. Human metabolic reconstruction is expected not only to become a prototype for other mammalian reconstructions but hopefully also to enable significant dimensions in the study of human systems biology. The future promise for individualized medicine and treatment will need a context to integrate and analyze data, and models resulting from these reconstructions can play a significant role in fulfilling this need. However, the development of cell-type or context-specific models will require the integration of various types of data, including transcriptomic, proteomic, fluxomic, and metabolomic measurements. Achieving these ambitious goals will require top–down data sets in conjunction with quantitative bottom-up reconstructions such as the one this study has tried to make.

References

1. International Human Gene Sequencing Consortium (2004). Nature 431: 931-945.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304-1351.
3. McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, et al. (2001) A physical map of the human genome. Nature 409: 934-941.
4. Reed JL, Famili I, Thiele I, Palsson BO (2006) Towards multidimensional genome annotation. Nat Rev Genet 7: 130-141.
5. Sheikh K, Forster J, Nielsen LK (2005) Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of Mus musculus. Biotechnol Prog 21: 112-121.
6. Fong SS, Burgard AP, Herring CD, Knight EM, Blattner FR, et al. (2005) In silico design and adaptive evolution of Escherichia coli for production of lactic acid. Biotechnol Bioeng 91: 643-648.
7. Fong SS, Joyce AR, Palsson BØ (2005) Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. Genome Res 15: 1365-1372.
8. Fong SS, Marciniak JY, Palsson BØ (2003) Description and interpretation of adaptive evolution of Escherichia coli K-12 MG1655 by using a genome-scale in silico metabolic model. J Bacteriol 185: 6400-6408.
9. Hua Q, Joyce AR, Fong SS, Palsson BØ (2006) Metabolic analysis of adaptive evolution for in silico-designed lactate-producing strains. Biotechnol Bioeng 95: 992-1002.
10. Ibarra RU, Edwards JS, Palsson BØ (2002) Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420: 186-189.
11. Reed JR, Patel TR, Chen KH, Joyce AR, Applebee MK, et al. (2006) Systems approach to refining genome annotation. Proc Natl Acad Sci USA 103: 17480-17484.
12. Natalie CD, Scott AB, Neema J (2007) Proc Natl Acad Sci USA 104: 1777-1782.
13. Pedro R, Jonathan W, Michelle LG (2004) Genome Biology 6: R2.