HinCyc: A Knowledge Base of the Complete Genome and Metabolic Pathways of *H. influenzae*

Peter D. Karp and **Christos Ouzounis** and **Suzanne Paley**
Artificial Intelligence Center
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025
{pkarp,ouzounis,paley}@ai.sri.com
v: 415-859-6375, f: 415-859-3735

Abstract[1]

We present a methodology for predicting the metabolic pathways of an organism from its genomic sequence by reference to a knowledge base of known metabolic pathways. We applied these techniques to the genome of *H. influenzae* by reference to the EcoCyc knowledge base to predict which of 81 metabolic pathways of *E. coli* are found in *H. influenzae*. The resulting prediction is a complex hypothesis that is presented in computer form as HinCyc: an electronic encyclopedia of the genes and metabolic pathways of *H. influenzae*. HinCyc connects the predicted genes, enzymes, enzyme-catalyzed reactions, and biochemical pathways in a WWW-accessible knowledge base to allow scientists to explore this complex hypothesis.

Introduction

The recent completion of the genomic sequence of *H. influenzae* (Fleishmann *et al.* 1995) presents an opportunity for predicting the metabolic map of this organism. We present a methodology for predicting the metabolic pathways of an organism from its genomic sequence by reference to a knowledge base (KB) of known metabolic pathways. Our approach uses the EcoCyc KB of *E. coli* pathways (Karp *et al.* 1996). The resulting prediction is a complex hypothesis that is presented in computer form as HinCyc: an electronic encyclopedia of the genes and metabolic pathways of *H. influenzae*.[2]

We created HinCyc by reusing the knowledge-base management and graphical-user interface tools that underly EcoCyc. HinCyc serves several purposes: It allows a researcher to examine our prediction of the metabolic complement of *H. influenzae*, and to explore other aspects of the *H. influenzae* genome in a user-friendly fashion (such as relationships among the DNA sequence, the genomic map, and gene-enzyme relationships). HinCyc also contains extensive cross-links to EcoCyc to facilitate the analysis of the *H. influenzae*

---

[1] This paper appears in the proceedings of the ISMB-96 conference.

[2] HinCyc is available via WWW URL http://www.ai.sri.com/ecocyc/hincyc.html.

genome relative to that of *E. coli*, thus allowing a researcher to view *H. influenzae* through an *E. coli* lens.

In brief, our methodology for creating HinCyc results from automating the following reasoning. For a given pathway defined in EcoCyc we can ask: Is there evidence that this pathway also occurs in *H. influenzae*? The primary evidence we use for the presence of a pathway is an estimate of whether *H. influenzae* can catalyze the reactions in the pathway. In turn, we consider there to be evidence for a reaction if there is evidence that an enzyme known to catalyze that reaction is found in *H. influenzae*.

Figure 1 shows the processing used to construct HinCyc. For every *H. influenzae* gene identified by The Institute for Genomic Research (TIGR) (Fleishmann *et al.* 1995), we created a gene object in the HinCyc KB. We then searched each *H. influenzae* protein sequence against the 3472 *E. coli* protein sequences in SWISS-PROT using BLAST, to identify their *E. coli* homologs. Next, we created a HinCyc polypeptide object for each *H. influenzae* gene product, and we linked each polypeptide to its *E. coli* homologs. For each HinCyc polypeptide whose strongest *E. coli* homolog is an enzyme in EcoCyc, we copied the reaction object(s) connected to the EcoCyc enzyme into HinCyc. We then asked for each metabolic-pathway frame defined in EcoCyc: how many of its reactions are catalyzed in *H. influenzae*? For those pathways for which sufficient evidence is found, we copied the EcoCyc pathway object into HinCyc.

Our approach omits some of the evidence that is relevant to evaluating the presence of a pathway P in an organism, such as evidence that other pathways that connect to P are also present. However, we consider it worthwhile to empirically evaluate what results can be obtained with this relatively simple approach to understand what is gained from more complex approaches. In future work we plan to determine the sensitivity of the prediction with respect to improvements to the reasoning process.

The parallel organization of EcoCyc and HinCyc is shown in Figure 2. They contain information about the genes, enzymes, reactions, pathways, and compounds of their respective species.

The remainder of this paper describes these steps

SwissProt
E. coli
subset

TIGR genes

EMBL/SRI
genes

Create
HinCyc
genes

Similarity
search to
E. coli

Link to
E. coli
homologs

Evaluate
pathways

Link
HinCyc
enzymes to
reactions

Import
EcoCyc
reactions,
pathways

Create
HinCyc
protein
complexes

Create
HinCyc
polypeptides
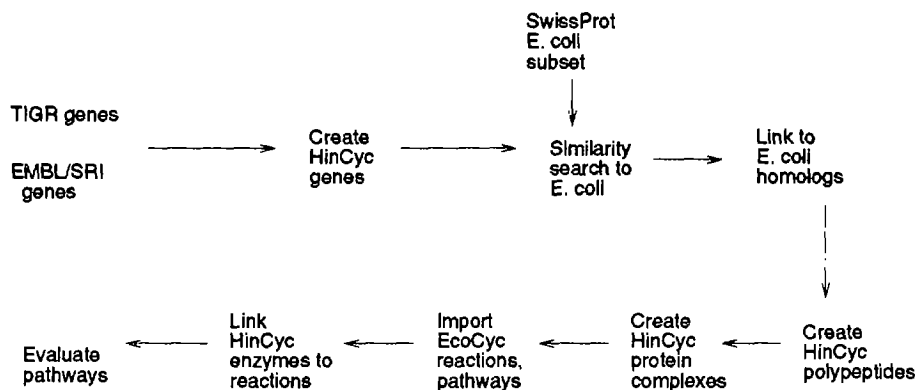
Figure 1: The flow of data and operations in the construction of HinCyc.

SwissProt   GenBank

HinCyc

Proteins	Genes	Pathways	Compounds	Reactions	EcoCyc Software

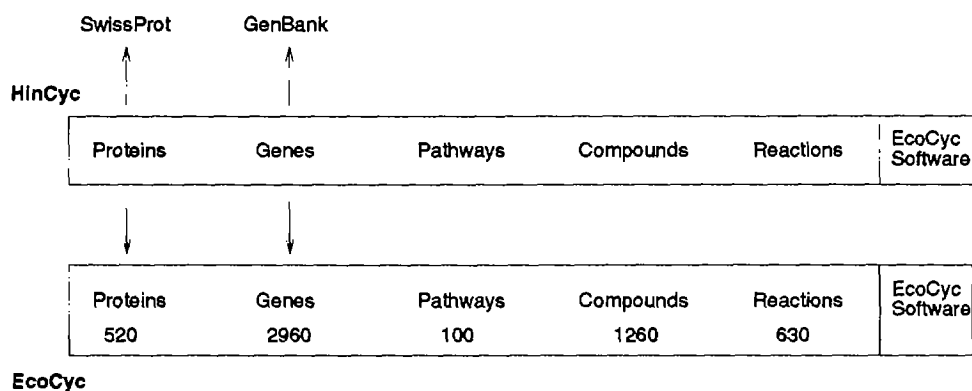| Proteins	Genes	Pathways	Compounds	Reactions	EcoCyc Software
520	2960	100	1260	630	

EcoCyc

Figure 2: HinCyc and EcoCyc encode the same classes of information about their respective organisms. Each HinCyc polypeptide is linked to all homologous EcoCyc polypeptides to facilitate analysis of *H. influenzae* with respect to *E. coli*. HinCyc genes and proteins are also linked to Genbank and to SWISS-PROT, respectively. The number of objects of each type within the EcoCyc KB are also shown. Note that reactions and compounds are not species specific, and therefore can be transferred from EcoCyc to HinCyc.

in more detail, and summarizes our prediction of the *H. influenzae* metabolic map. We also discuss the assumptions that our method relies on, to provide an understanding of the uncertainties in the prediction.

Background

The Encyclopedia of *E. coli* Genes and Metabolism (EcoCyc) is a large KB that describes the genome and the metabolic pathways of *E. coli* (Karp *et al.* 1996; Karp & Paley 1996). The information in EcoCyc was gleaned from the biomedical literature and from existing databases. The KB currently contains over three-quarters of the known metabolic pathways of *E. coli*. EcoCyc describes the reactions in each pathway, and the enzyme that carries out each reaction, including its cofactors, activators, inhibitors, and the subunit structure of the enzyme. The genes encoding the subunits of an enzyme are listed, as is the map position of most genes on the *E. coli* chromosome.

The KB is managed using HyperTHEO — a Frame Knowledge Representation System (FRS) (Karp & Paley 1996). The EcoCyc program includes a graphical user interface that generates displays of genes, enzymes, reactions, pathways, and compounds. It performs automatic layout of metabolic pathways, and includes a genomic map browser. The EcoCyc GUI runs as both an X-Windows application. and as a WWW server at http://www.ai.sri.com/ecocyc/browser.html. In constructing HinCyc we reused most of the software infrastructure that supports EcoCyc, including the EcoCyc schema (Karp & Riley 1993; Karp & Paley 1994), HyperTHEO, and the EcoCyc GUI.

Genes

We created gene objects in HinCyc for the 1743 *H. influenzae* genes identified by TIGR (Fleishmann *et al.* 1995).[3] Additional function predictions were added as a result of analysis by Casari et al. (Casari *et al.* 1995). Frameshifts identified by continuing analyses led us to merge several open reading frames (ORFs) and omit 34 of the original ORFs. We follow updates to the sequence annotations as they become available.

Each gene object contains the location of the start and stop codons within the complete sequence, the direction of transcription, the gene name (all of the preceding were assigned by TIGR), and a comment describing the source of the gene (from Fleischmann et al. or Casari et al.). Each gene is linked to the Genbank entry that contains it, and HinCyc contains the complete *H. influenzae* DNA sequence as determined by Fleischmann et al. (Fleishmann *et al.* 1995).

All *H. influenzae* ORFs were searched against the collection of *E. coli* proteins available from SWISS-PROT version 32 (Bairoch & Boeckmann 1994), using BLAST (Altschul *et al.* 1990). ORFs were corrected for composition bias (unpublished). Matches

---

[3] We downloaded a file of gene identifications from www.tigr.org on November 1, 1995.

with probability values less than 10^{-6} were considered as significant. Sequence-similarity values for pairs of homologous sequences were obtained.

Proteins

HinCyc proteins are inferred in two steps: first polypeptides, then protein complexes. A polypeptide object is created for each *H. influenzae* gene that has a protein product — either as given by the *H. influenzae* section of SWISS-PROT or because a gene has homology to an *E. coli* protein. The HinCyc polypeptide object is linked to the gene that encodes it, to its *E. coli* homologs (in both EcoCyc and SWISS-PROT), and to its SWISS-PROT entry (if available).

This paper presents a conservative prediction of HinCyc pathways, because at several processing steps, we removed inferences that might be inaccurate in order to arrive at a relatively minimal but more reliable prediction. For example, in this phase we removed links from a HinCyc polypeptide to an EcoCyc polypeptide P_1 in the case where a significantly stronger homolog P_2 existed in *E. coli* (where P_1 has a p-value at least 10^{20} times that of P_2), but P_2 was not present in EcoCyc. This step removed our functional assignments for 26 *H. influenzae* proteins.

The second step is to create protein-complex objects. EcoCyc contains extensive information about the quaternary structures of *E. coli* enzymes, and about the enzymatic activities of the multimeric form of an enzyme versus the activities of its monomers. For example, EcoCyc records that the product of the *lysC* gene (2.7.2.4) acts only as a dimer. In contrast. the products of the genes *trpE* and *trpD* (to choose a complex example) catalyze the reaction 4.1.3.27 in the form $TrpE_2TrpD_2$, and catalyze the reaction 2.4.2.18 in the form $TrpD_2$.

An accurate prediction of the catalytic capabilities of a given cell must be based on an accurate prediction of what polypeptides are found in the cell, and of what protein complexes those polypeptides form. There are of course many cases involving homomultimers where prediction is more straightforward. For example, in *E. coli* the product of *pykA*, pyruvate kinase II (2.7.1.40), acts as a homotetramer. Since *H. influenzae* contains a homolog for this gene, it seems reasonable to assume that *H. influenzae* catalyzes this reaction. But what if we observed that *H. influenzae* contained a homolog to *E. coli trpD*, but no homolog to *trpE*? If we find any component of a protein complex, should we assume that all activities of all members of the complex are present (e.g., both 4.1.3.27 and 2.4.2.18)? Or should we assume that only those activities known to be specific to the polypeptides for which evidence exists are present (e.g., only 2.4.2.18)? Either line of reasoning might be correct: the gene could be missing from *H. influenzae* altogether, or, less likely, the *H. influenzae trpE* could have diverged beyond recognition from the *E. coli trpE*.

This hypothetical *trpE* example is relatively easy to resolve, because both of the enzyme activities in ques-

tion occur in the biosynthetic pathway for tryptophan. It seems unlikely that the cell would retain one reaction in the pathway and not the other. The situation becomes murkier if the reaction in question (2.4.2.18) is also found in another pathway. This particular reaction is only used in tryptophan biosynthesis, but there are 45 reactions in EcoCyc that are used in more than one pathway.

Consider a more extreme (and rare) example. The product of the *E. coli lpd* gene is found in three different protein complexes (the 2-oxoglutarate dehydrogenase complex, the gcv system, and the pyruvate dehydrogenase multienzyme complex). These enzymes carry out five different reactions in three different pathways. The *lpd* gene does have a homolog in *H. influenzae*; if we observe only that homolog, but homologs for none of the other half-dozen subunits of these protein complexes, which reaction(s) and pathway(s) would we claim are present?

The succinate dehydrogenase enzyme (1.3.99.1) in *E. coli* provides another interesting case. Two of its subunits are catalytic and two are membrane proteins. Imagine that in *H. influenzae* we found homologs for the membrane proteins only: should we assume that *H. influenzae* can catalyze 1.3.99.1, or is it more likely that the membrane subunits became coopted for another use? (The actual situation is the exact opposite: *H. influenzae* contains homologs for the catalytic subunits but not for the membrane subunits!) In any event, to support this type of reasoning, EcoCyc should explicitly encode which subunits are responsible for each activity of an enzyme complex, which it currently does not (it does encode observations that a given subunit can catalyze some reaction *independent* of a complex of which it is normally part).

The actual algorithm we use to construct protein complexes in HinCyc by analogy to protein complexes in EcoCyc is as follows. Figure 3 shows a complicated protein complex in EcoCyc where two copies each of subunits A and B form a complex X, which aggregates with a single copy of the homotetramer Y, to form Z (a number of such complexes of complexes are found in *E. coli*). Imagine that *H. influenzae* contains polypeptides A' and C' as subunits. As a minimal commitment about the actual situation in *H. influenzae*, we create only the complex Z'. We do not create intermediate complexes X' nor Y', nor do we create a hypothetical B', nor do we fill in coefficients for A' or C' — all on the grounds that we do not have enough knowledge about the conservation of subunit relationships across species to make such detailed inferences.

HinCyc contains 1611 polypeptides, of which 1236 have at least one *E. coli* homolog. 277 of the polypeptides are homologous to an EcoCyc polypeptide (EcoCyc describes only proteins that are enzymes). The 277 *H. influenzae* polypeptides show significant homology to 313 EcoCyc polypeptides, but if we consider only the highest-scoring *E. coli* homolog for each *H. influenzae* polypeptide, we find 256 EcoCyc polypeptides, meaning that a given EcoCyc polypeptide is

sometimes the homolog for more than one *H. influenzae* polypeptide.

Of the 277 HinCyc polypeptides, 115 are monomers, and 162 are grouped into 138 protein complexes; 108 of the latter are properly homomultimers. Of the 30 heteromultimers, 21 have no missing monomers, 6 have one missing monomer, and 3 have more than one missing monomer: formate dehydrogenase-0 (1.2.1.2) has 1 of 3 subunits present, NADH dehydrogenase I (1.6.5.3) has 1 of 14 subunits, and hydrogenase-3 (1.12.1.2) has 1 of 5 subunits. These are all curious cases since the majority of the subunits are missing. These three complexes were deleted in favor of a more conservative prediction.

Reactions

The next step is to determine what reactions are catalyzed by each *H. influenzae* enzyme. The inference we applied is: For a HinCyc enzyme complex E', let E be the EcoCyc homolog of E'. Let S be the set consisting of E plus all subunits of E'. Let R be the union over S of all reactions catalyzed by each element of S. Infer that all reactions in R are catalyzed by E'.

Note that as discussed in Section , this inference is subject to several possible errors. If not all subunits of the *E. coli* protein complex are found in *H. influenzae*, the *H. influenzae* enzyme may not have all the activities of its *E. coli* counterpart. Unfortunately, EcoCyc does not specify which subunits of an enzyme complex are responsible for which activities of the complex. Another potential source of error is the fact that the preceding functional assignment step actually ignores the functional assignments made by TIGR and by SWISS-PROT. The TIGR assignments are extremely difficult to utilize in a computational environment since they are purely textual and omit EC numbers. The SWISS-PROT entries, however, include EC numbers. We therefore compared the EC numbers that we inferred using the sequence-similarity search against *E. coli* proteins, to the EC numbers in SWISS-PROT.

Of all genes identified by TIGR, SWISS-PROT has assigned 439 EC numbers to the polypeptide products of 429 genes (there are roughly 10 multifunctional enzymes in *H. influenzae*). For 213 of the 429 HinCyc polypeptides, their EC numbers existed in EcoCyc, and EcoCyc describes the enzyme associated with that EC number. The remainder represent either reactions that do not occur in *E. coli*, incomplete information in EcoCyc, or partially specified EC numbers (for example, numbers such as 1.2.3.- are used to informally indicate that the class of the reaction is known but the enzyme committee has not yet assigned an official number).

Using BLAST, we identified 261 of the 1519 genes that have homologous polypeptides in EcoCyc. We actually identified many more *E. coli* homologs, but the polypeptides were not present in EcoCyc, either because they are not involved in central metabolism or because EcoCyc is incomplete. Of the 261 Hin-
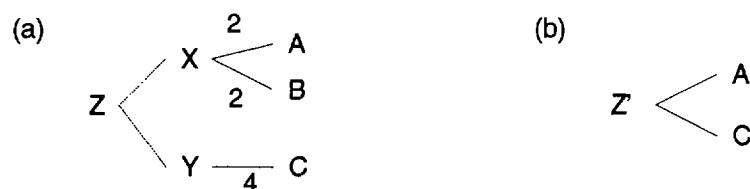
Figure 3: (a) A multienzyme complex in *E. coli*; (b) the analogous complex in *H. influenzae*.

Cyc polypeptides, 225 of them have EcoCyc homologs that have associated EC numbers. We can therefore compare the EC numbers that would be inferred for these 225 HinCyc polypeptides with the EC numbers supplied by SWISS-PROT for the preceding 213 HinCyc polypeptides. The intersection of these two sets of polypeptides yields 201 polypeptides. The EC numbers matched in 193 cases. Of the eight conflicting cases, one shows a match between the TIGR and SWISS-PROT functional assignment and our second choice *E. coli* homolog. Two of the cases involve partial EC numbers that match, and we found that the textual descriptions also match. The remaining five cases are complete mismatches, and it remains to be determined which functional assignment is the correct one. There could be other conflicts between our functional assignments and those of TIGR and SWISS-PROT, but these cannot be determined automatically.

In summary, our functional assignments are consistent with the SWISS-PROT assignments. The advantage of not simply using the SWISS-PROT assignments is that we are able to infer additional functions for enzymes whose function is not described in SWISS-PROT, or whose function has not yet been given an EC number, but is contained in EcoCyc.

It is also interesting to determine which reactions are present in *H. influenzae*, but apparently do not occur in *E. coli*. We computed a list of all *H. influenzae* genes to which SWISS-PROT has assigned complete EC numbers, but which are not present in *E. coli* (as determined by EcoCyc). This procedure resulted in a list of seven EC numbers that are present in *H. influenzae* for which there is no evidence of occurrence in *E. coli*. They are: 2.3.1.31, 2.7.2.2, 3.1.21.4, 3.4.21.72, 3.4.24.57, 4.1.3.6, and 6.2.1.22. The accuracy of this result is limited by the fact that EcoCyc does not yet contain information about all enzymes that are known to occur in *E. coli*, nor is the full set of enzymes that occur in *E. coli* known.

Pathways

Our pathway predictions for *H. influenzae* are shown in Tables 1, 2 and 3, grouped roughly by the degree of evidence for the pathways. Table 1 shows those *E. coli* pathways whose presence in *H. influenzae* is likely since almost all of the reactions for these pathways are catalyzed by *H. influenzae*. Each row in these tables gives a pathway name followed by two columns of evidence. Column 1 lists the (unreduced) fraction of those

reactions in the pathway for which an enzyme has been inferred to occur in *H. influenzae*. For example, in the first line of Table 2, eight of the thirteen reactions of the *E. coli* fermentation pathway can be catalyzed by *H. influenzae*.

The third column provides evidence *against* the presence of a pathway. Of the eight reactions of fermentation that *H. influenzae* can catalyze (Table 2, line 1), five of those reactions are also used in four other pathways. This evidence could support an argument that fermentation does not occur in *H. influenzae*, since only three of the fermentation reactions observed in *H. influenzae* are unique to that pathway — their presence cannot be explained by participation in a different pathway. Column 3 is more relevant in explaining away those pathways for which only marginal evidence exists. Consider the TCA cycle in Table 2: although we have evidence that *H. influenzae* can catalyze four of its nine reactions, two of those reactions (1.3.99.1 and 1.1.1.37) are used in two different pathways (note that the number of reactions (numerator) in the second column is always computed relative to the reactions in the (numerator of) the first column — those for which there is evidence that they occur in *H. influenzae*). Therefore, only two unique TCA cycle reactions are found in *H. influenzae*, which may be remnants of a lost TCA cycle. On the other hand of course, the remaining four enzymes (which catalyze five reactions) may be discovered at a later time.

Overall, *H. influenzae* has a fairly complete set of biosynthetic pathways for amino acids, nucleotides, and fatty acids. Surprisingly, a number of cofactor-biosynthesis pathways appear to be missing from *H. influenzae*. Also missing are several catabolic pathways for carbohydrates, which could be linked to the growth-medium versatility of this organism.

Table 3 should be considered the least solid prediction; because we used more conservative evidence in making these predictions, a *lack of evidence* for a given pathway should not be considered strong evidence that the pathway *does not occur*. That is, were we performing a more liberal prediction, some evidence would no doubt appear for some of these pathways.

Limitations

The pathway predictions computed by our methods should be viewed as tentative hypotheses that are likely to contain errors, and that will undergo revision over time as better input data and better prediction meth-

Pathway	Evidence	Other pwys
purine biosynthesis	12/12	2/1
pyrimidine ribonucleotide/ribonucleoside metabolism	11/11	3/2
histidine biosynthesis	10/10	0/0
threonine biosynthesis	5 /5	3/1
valine biosynthesis	5 /5	2/2
fatty acid elongation, unsaturated	4 /4	3/1
deoxyribonucleotide metabolism	4 /4	2/1
leucine biosynthesis	4 /4	0/0
pyruvate dehydrogenase	3 /3	0/0
serine biosynthesis	3 /3	1/1
glycogen biosynthesis	3 /3	0/0
peptidoglycan precursor biosynthesis	3 /3	2/2
phosphatidic acid synthesis	2 /2	0/0
glucosamine catabolism	2 /2	0/0
glycine biosynthesis	2 /2	2/2
cysteine biosynthesis	2 /2	0/0
removal of superoxide radicals	2 /2	0/0
aspartate biosynthesis	1 /1	0/0
glutamine biosynthesis	1 /1	0/0
peptidoglycan biosynthesis	8 /9	1/1
glycolysis	8 /9	6/2
biosynthesis of lysine and diaminopimelate	8 /9	2/1
fatty acid biosynthesis, initial steps	7 /8	0/0
methionine biosynthesis	7 /8	5/3
menaquinone biosynthesis	6 /7	0/0
fatty acid oxidation pathway	6 /7	0/0
non-oxidative branch of the pentose phosphate pathway	5 /6	3/2
formylTHF biosynthesis	13/16	10/5
glycerol metabolism	4 /5	0/0
fatty acid elongation, saturated	4 /5	3/1
isoleucine biosynthesis	4 /5	0/0
nucleotide metabolism	22/29	4/2
deoxypyrimidine nucleotide/side metabolism	9 /12	3/2
lipid-A-precursor biosynthesis	6 /8	2/1
proline biosynthesis	3 /4	0/0
folic acid biosynthesis	11/15	6/2
gluconeogenesis	8 /12	6/2
riboflavin, FMN and FAD biosynthesis	6 /9	0/0
oxidative branch of the pentose phosphate pathway	2 /3	1/1
ppGpp metabolism	2 /3	0/0
methyl-donor molecule biosynthesis	2 /3	2/1
alanine biosynthesis	2 /3	2/2
tyrosine biosynthesis	2 /3	1/1
phenylalanine biosynthesis	2 /3	1/1
pyruvate oxidation pathway	2 /3	2/1
ribose catabolism	2 /3	2/2

Table 1: We find strong evidence that these *E. coli* pathways occur in *H. influenzae*. To determine exatly what is meant by a given pathway name, find that pathway through the EcoCyc WWW server (see Section 2).

Pathway	Evidence	Other pwys
fermentation	8 /13	5/4
glucose 1-phosphate metabolism	3 /5	2/1
tryptophan biosynthesis	3 /5	0/0
KDO biosynthesis	3 /5	2/1
glyoxylate degradation	3 /5	1/1
(deoxy)ribose phosphate metabolism	5 /9	4/3
pyrimidine biosynthesis	3 /6	1/1
Entner-Doudoroff pathway	1 /2	0/0
4-aminobutyrate degradation	1 /2	0/0
asparagine biosynthesis	1 /2	0/0
L-alanine degradation	1 /2	1/1
glycine cleavage	1 /2	1/1
TCA cycle	4 /9	2/2
arginine biosynthesis	3 /8	0/0

Table 2: We find moderate evidence that the these *E. coli* pathways occur in *H. influenzae*.

ods become available. Our method is limited by the following factors.

The quality of the input data. Incorrect identification of *H. influenzae* ORFs, or incorrect functional assignments could lead to false-negative or false-positive pathway predictions. For example, the incorrect identification of a given *H. influenzae* enzyme would provide incorrect evidence for the pathway containing that enzyme.

Incomplete datasets. Many *H. influenzae* genes do not have functional assignments, the sequence of *E. coli* has not been completed, and EcoCyc does not describe a number of *E. coli* enzymes and pathways (both known and unknown). These factors would cause false-negative predictions. For example, if a given enzyme had been sequenced in *H. influenzae* but not in *E. coli*, the *H. influenzae* enzyme would not be given a *E. coli* homolog, would not have a reaction inferred for it, and therefore would not provide evidence that its pathway occurs in *H. influenzae*.

Pathways unique to *H. influenzae*. Our method makes no prediction regarding pathways that exist in *H. influenzae* but not in *E. coli* since our available pathway KB spans *E. coli* only. We do, however, identify reactions that are catalyzed in *H. influenzae* but not *E. coli*. This limitation would be more significant when analyzing organisms with a greater evolutionary distance from *E. coli*, in which we would expect to find a larger number of pathways not found in *E. coli* — although just how great the metabolic variation is among different organisms is an open question. These novel pathways could be of two forms: novel *implementations* (e.g., although *E. coli* can synthesize methionine, another organism might use a different pathway to synthesize methionine), and novel functions (e.g., a biosynthetic pathway to synthesize a compound that *E. coli* is unable to synthesize by any means).

It is not always possible to evaluate metabolic reconstructions from whole genomes. The metabolic pathways for *H. influenzae* are not well studied experimentally. The pathways of *M. genitalium*, whose genome has also been sequenced, are better understood, but are much simpler and therefore less interesting to predict because of the parasitic lifestyle of mycoplasmas.

Related Work

Gaasterland and Selkov defined the problem of metabolic prediction from sequence data one year ago in a comprehensive paper (Gaasterland & Selkov 1995). They considered a broad range of evidence and reasoning strategies that bear on this problem, such as considering experimentally determined growth-medium requirements of the organism, introducing default pathways from evolutionarily related organisms, and introducing new pathways to produce or consume "dangling compounds" in existing pathways.

Our method does not consider many of these sources of evidence, although we do consider a number of complications that they do not address, such as multimeric enzymes, reactions that occur in multiple pathways, and the types of evidence that are supplied by sequencing projects. Although our method uses less evidence, it is still able to make comprehensive predictions.

We do question the practice of introducing default pathways during the interpretation process, since it is unclear from how remote a taxon it is valid to retrieve a default pathway from, that is, how can we rule out postulating by default that every known metabolic pathway occurs in a given organism? It would be useful to gather data regarding the prior probability of every metabolic pathway in different regions of the evolutionary tree, to guide the introduction of default pathways.

Discussion

A number of issues arose in the course of this project that, if resolved, would make similar analyses easier in the future.

Several questions arise regarding enzyme complexes: To what degree are quaternary protein structures conserved through evolution? If several polypeptides are subunits of a complex in one species, how likely are homologs of those polypeptides to be associated in a

Pathway	Evidence	Other pwys
UDP-N-acetylglucosamine biosynthesis	2 /6	1/1
NAD phosphorylation and dephosphorylation	1 /3	0/0
dTDP-rhamnose biosynthesis	1 /4	0/0
galactose, galactoside and glucose catabolism	2 /10	2/1
pyridoxal 5'-phosphate biosynthesis	1 /5	0/0
glyoxylate cycle	1 /6	1/1
pyridine nucleotide synthesis	1 /6	0/0
polyamine biosynthesis	1 /6	0/0
biosynthesis of proto- and siroheme	2 /13	0/0
thiamin biosynthesis	1 /9	0/0
pantothenate and coenzyme A biosynthesis	1 /10	0/0
glycogen catabolism	0 /7	0/0
glycolate metabolism	0 /6	0/0
D-glucuronate catabolism	0 /5	0/0
fucose catabolism	0 /5	0/0
rhamnose catabolism	0 /5	0/0
sulfate assimilation pathway	0 /4	0/0
D-galacturonate catabolism	0 /4	0/0
propionate metabolism, methylmalonyl pathway	0 /3	0/0
degradation of short-chain fatty acids	0 /2	0/0
glutamate biosynthesis	0 /1	0/0

Table 3: We find little or no evidence that these *E. coli* pathways occur in *H. influenzae*.

related species? How does enzyme activity vary as a function of changes in quaternary structure?

EC numbers proved to be very useful in this project because they allowed automatic cross-checking of the functions of our computed *E. coli* homologs with the functional assignments in SWISS-PROT. We strongly suggest that future sequencing projects supply EC numbers with their functional predictions. We also note that the EC system is currently quite incomplete: EcoCyc describes 87 enzymes whose activities do not yet have an EC number (about 20% of the enzymes in EcoCyc). Furthermore, it would be useful to have an EC-like system for other types of proteins besides enzymes (e.g., transport proteins), to facilitate reasoning about their functions.

Enzyme variability complicates metabolic prediction in several ways. For example, if *H. influenzae* employs an enzyme E to catalyze a reaction R that is catalyzed by an enzyme E' in *E. coli* but E has no sequence similarity to E' (or to any other known enzyme that catalyzes R), how can we recognize that *H. influenzae* can catalyze R? Variability in substrate specificity can also introduce false positive or negative predictions: even if E does have sequence similarity to E', E might have evolved to carry out reaction R' instead of R, or to carry out reaction R'' in *addition* to R, both of which cases are difficult to predict.

This project raises a host of fundamental questions about the process of predicting gene function by running similarity searches in sequence databases. Those functional predictions are the principal input of our metabolic-prediction techniques, and of those of Selkov and Gaasterland, and therefore these methods are intimately dependent on these predictions. The following questions arise.

How can we assess the reliability of a functional prediction based on a similarity search? The p-value reported by a program such as BLAST is the probability that a similarity between two sequences arose by chance. It is not the probability that the two sequences share the same function! The relationship between these probabilities will vary from one protein family to another, and across species, but it would be useful in deciding what predictions are most likely to be correct.

Why is it that the functional predictions that are published by many of the sequencing projects do not contain estimates of reliability? Even if programs such as BLAST cannot compute such estimates automatically, the scientists who make the final assignment after looking at multiple matching sequences, alignments, profile searches, and so forth, are in a good position to make at least a rough estimate of its reliability.

How can we resolve conflicting functional predictions for genes? Given that there is uncertainty in functional assignments, why not publish several alternative predictions for a gene if various functions are reasonably likely? Such predictions would be useful to our method. Imagine if our method were told that a given gene could catalyze either reaction R_1 or R_2, where R_1 is found in pathway P_1 and R_2 is found in P_2. If there were much more other evidence for the presence of P_2 than P_1, we might decide that R_2 is the more likely interpretation. In contrast, we might have assigned greater likelihood to the presence of P_1, had the original functional assignment been R_1 only.

To what degree does absence of a similarity match imply absence of a function in a complete genome? Fleischmann et al. found that no *H. influenzae* genes match any of the known sequences for four of the en-

zymes in the TCA cycle (Fleishmann *et al.* 1995). If we could transform the lack of a similarity match into a probability that these enzymes were truly missing from *H. influenzae* (as opposed to being undetected either because of sequence divergence, or independent evolution of the same function), we could give a better estimate that the TCA cycle itself is absent. Such an estimate might be based on the number of known sequences of a given enzyme, their sequence variation as a function of evolutionary distance, and the evolutionary distance between the organism of interest and the closest known sequence.

Summary

We have presented techniques for predicting the metabolic map of an organism from its predicted gene products. We used those techniques to predict the metabolic map of *H. influenzae* by analogy to the metabolic map of *E. coli*. Out of 81 metabolic pathways that occur in *E. coli*, we find strong evidence for the presence of 46 of the pathways in *H. influenzae*; we find medium evidence for the presence of 14 pathways, and we find little or no evidence for 21 pathways. Our prediction is presented as HinCyc, a KB that links the predicted genes of *H. influenzae* to the predicted products of those genes; enzyme products are linked to the reactions that they are predicted to catalyzed, which in turn are linked to the pathways that contain them. Scientists can employ HinCyc to examine gene–function relationships in *H. influenzae*, and can further analyze *H. influenzae* relative to *E. coli* because of extensive links between HinCyc and EcoCyc.

Acknowledgments

References

Altschul, S.; Gish, W.; Miller, W.; Myers, E.; and Lipman, D. 1990. Basic local alignment search tool. *J Mol Bio* 215:403–410.

Bairoch, A., and Boeckmann, B. 1994. The SWISS-PROT protein sequence data bank: current status. *Nucleic Acids Research* 22:3578–3580.

Casari, G.; Andrade, A.; Bork, P.; Boyle, J.; Daruvar. A.; Ouzounis, C.; Schneider, R.; Tamames, J.; Valencia, A.; and Sander, C. 1995. Challenging times for bioinformatics. *Nature* 376:647–648.

Fleishmann, R.; Adams, M.; White, O.; Clayton, R.; Kirkness, E.; Kerlavage, A.; Bult, C.; Tomb, J.-F.; Dougherty, B.; Merrick, J.; McKenney, K.; Sutton, G.; FitzHugh, W.; Fields, C.; Gocayne, J.; Scott, J.; Shirley, R.; Liu, L.-I.; Glodek, A.; Kelley, J.; Weidman, J.; Phillips, C.; Spriggs, T.; Hedblom, E.; Cotton, M.; Utterback, T.; Hanna, M.; Nguyen, D.; Saudek, D.; Brandon, R.; Fine, L.; Fritchman, J.; Fuhrmann, J.; Geoghagen, N.; Gnehm, C.; McDonald, L.; Small, K.; Fraser, C.; Smith, H.; and Venter, J. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae. Science* 269:496–512.

Gaasterland, T., and Selkov, E. 1995. Reconstruction of metabolic networks using incomplete information. In Rawlings, C.; Clark, D.; Altman, R.; Hunter, L.; Lengauer, T.; and Wodak, S., eds., *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, 127–135. Menlo Park, CA: AAAI Press.

Karp, P., and Paley, S. 1994. Representations of metabolic knowledge: Pathways. In Altman, R.; Brutlag. D.; Karp, P.; Lathrop, R.; and Searls, D., eds., *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 203–211. Menlo Park, CA: AAAI Press.

Karp, P., and Paley, S. 1996. Integrated access to metabolic and genomic data. *Journal of Computational Biology* 3(1):191–212.

Karp, P., and Riley, M. 1993. Representations of metabolic knowledge. In Hunter, L.; Searls, D.; and Shavlik, J., eds., *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, 207–215. Menlo Park, CA: AAAI Press.

Karp, P.; Riley, M.; Paley, S.; and Pellegrini-Toole, A. 1996. EcoCyc: Electronic encyclopedia of *E. coli* genes and metabolism. *Nuc. Acids Res.* 24(1):32–40.