

Challenging times for bioinformatics

SIR—On July 28, 1995, the timely public release of the complete DNA sequence of *Haemophilus influenzae* Rd made the world of biology richer, in one day, by 1,830,137 characters of genetic information encoding a complete bacterial construction kit of 1,743 predicted protein genes¹. While marveling at the advances in sequencing technology one wonders how biologists will cope with the onslaught of data from this and other genome projects. To illustrate recent advances in informatics technology, we have analyzed the open reading frames of *H. influenzae* with the aim of adding as

much biological knowledge as possible to each predicted protein sequence. The computer runs were completed in three days using the GeneQuiz software system² which identifies likely protein functions from amino acid sequence information in a completely automated fashion. This resulted in a new set of 148 putative protein functions in *H. influenzae* (see Table) and confirmed most of the 1,007 already identified by TIGR¹.

A thorough and reliable analysis of several thousand protein sequences in one week requires more efficient tools than are currently used for routine data-

base searches. Our approach to high efficiency biosequence informatics mirrors that of high-throughput sequencing: to start with existing technology, add robotics, and focus on solutions for rate-limiting steps. In computerized sequence analysis, the analogue of a robot is a program which embodies key rules learned from a human expert and applies these at high speed. The GeneQuiz system was developed during the analysis of yeast chromosomes and mycoplasma sequences³. It combines such an expert module with existing methods, such as programs for similarity searches in sequence databas-

New protein functions in *H. influenzae* identified by automated sequence analysis.

0017	Formate acetyltransferase 1	0572	3D	PmpA peroxisomal membrane	1090	HelB. Cytochrome c biogenesis
0052	C4-dicarboxylate-binding periplasmic		#m	protein + Glutharedoxin	1091	HelC. Cytochrome c biogenesis
0053	3D RspB. Starvation sensing	0574	3D	Peptidyl-prolyl <i>cis-trans</i> isomerase	1092	Cytochrome c biogenesis
0121	UDP-N-acetylmuramate alanine ligase	0597		Hydrolase	1093	Pyoverdine transport. Cytochrome c biogenesis
0126	ABC transporter	0624	#m	Fmv+Fmu ATP-binding protein		
0131	Major ferric iron binding	0626		Large conductance mechanosensitive channel	1094	Ccl1/Ycf5. Cytochrome c biogenesis
0135	AraJ. Processing of arabinose polymers				1095	HelX. Cytochrome c biogenesis
0136	ThdF. Possible thiophene and furan oxidation	0634		NifH3. Nitrogen fixation	1104	Sialic acid transporter
0146	C4-dicarboxylate-binding periplasmic	0647		Mg(2+) transport ATPase C	1125	Transaldolase B
0164	#f NADH:ubiquinone oxidoreductase subunit A	0658		ABC transporter	1147	IIA-ntr. Nitrogen fixation.
0167	NADH:ubiquinone oxidoreductase	0718		VacJ. Intercellular spreading	1148	ABC transporter
0168	#f NADH:ubiquinone oxidoreductase	0723		TrkH. Potassium uptake	1153	3D Hypoxanthine phosphoribosyltransferase
0182	Glucokinase	0731	#f	Phenoxazinone synthase	1155	Activator of anaerobic ribonucleoside-triphosphate reductase
0183	Na(+)-linked D-alanine glycine permease	0740	3D	Phosphomannomutase		Thioredoxin
0184	Esterase D	0756		NlpD. Cell wall degradation	1165	Phosphate acetyltransferase
0258	#f Glycosyl transferase	0773		3-oxoadipate CoA-transferase subunit B	1203	ABC transporter
0270	NifH3. Nitrogen fixation	0831		Cell wall formation	1252	IS150 transposase
0281	Proline/betaine transporter PPII	0837		CpxR. Transcriptional regulator	1280	Nap57/cbf5/yhba family
0291	Mercuric transport. Periplasmic component	0852		Multidrug resistance protein B	1289	ABC transporter
0299	Prepilin peptidase dependent protein D	0858		Methenyl-THF synthetase	1300	Carbonic anhydrase
0323	Lactoylglutathione lyase	0860		SpoU. rRNA methylase	1301	Ferredoxin
0333	tRNA (uracil-5)-methyltransferase	0864	3D	Elongation factor G	1309	3D Sodium/myo-inositol cotransporter
0336	Mog. Molybdopterin biosynthesis	0881		Octaprenyl-diphosphate synthase	1315	Phosphomannomutase
0345	3D YojB. 4Fe-4S iron sulphur centre, electron transfer	0906		Nitrogen fixation	1337	3D ABC transporter
0346	MauN. 4Fe-4S iron sulphur centre, electron transfer	0929		Glutathionylspermidine synthetase	1342	Bacterioferritin
0347	Cytochrome c	0934		NrfF. Cytochrome c biogenesis	1349	Transcriptional regulator
0350	AmpG. Permease in beta-lactamase induction	0935		HelX. Cytochrome c biogenesis	1364	Acyl carrier phosphodiesterase
0364	Penicillin-binding protein 5	0936		Ccl1. Cytochrome c biogenesis	1366	Zinc proteinase
0376	HesB. Nitrogen fixation cluster	0948		Virulence-associated protein B	1437	Exodeoxyribonuclease small subunit
0388	O-sialoglycoprotein endopeptidase	0955		Ttk. Transcriptional regulator	1439	Transketolase
0389	Slp. Outer membrane	0959		Chitobiase	1454	CycZ. Cytochrome biogenesis
0393	GTP-binding	0961		Histidine triad zinc binder	1463	3D Phosphomannomutase
0409	TagE. Cell wall degradation	0963		FAD synthetase	1476	Transcription repressor
0412	Drap deaminase	0964		MviN. Virulence factor	1546	DNA repair
0415	Thiamine biosynthesis	0965		30s ribosomal protein S20	1555	FtsX. Cell division
0417	Thiamine biosynthesis	0977		Cell filamentation	1590	DNA polymerase III subunits γ and τ
0418	Proline/betaine transporter PPII	0979		NifH3. Nitrogen regulation	1607	HemM. Glutamyl-tRNA dehydrogenase
0456	ATP-binding pyrimidine kinase	0988	3D	3-Isopropylmalate dehydratase	1648	Amidotransferase
0464	Ribose 5-phosphate isomerase A	1001		Sporulation	1652	Extracellular endopeptidase
0469	Histidinol dehydrogenase	1004		Peptidyl-prolyl <i>cis-trans</i> isomerase	1665	Pepsinogen A-4
0493	IS150 transposase	1007		LytB. Penicillin tolerance	1684	3D Polyferredoxin
0494	NapA. Acid phosphatase	1010		3-hydroxyisobutyrate dehydrogenase	1688	NADH:Ubiquinone oxidoreductase
0508	MenG. Menaquinone biosynthesis	1019		Thiamine-binding. Periplasmic	1693	ModA. Molybdate-binding periplasmic
0509	MenA. Menaquinone biosynthesis	1020		Transport system permease	1694	Molybdenum transport
0519	NupC. Nucleoside permease.	1021		ABC transporter	1695	N-acetylmannosamine transferase
0520	Activator of pyruvate formate-lyase 1	1027	3D	L-Xylulose Kinase	1696	Glycosyltransferase in succinoglycan biosynthesis
0565	Phosphoglycolate phosphatase, chromosomal	1028		C4-dicarboxylate-binding. Periplasmic		
		1031		Malate dehydrogenase	1698	Glucosyltransferase in lipopolysaccharide synthesis
		1032		IclR family of transcriptional regulators	1721	IS904 transposase
		1043	3D	Ferredoxin	1723	Nitrogen fixation cluster, ATP-binding
		1070	#f	HrpA. ATP-dependent helicase	1730	Bifunctional urea amidolyase
		1089		ABC transporter		

The 1680 protein sequences were provided as of July 31, 1995, by the TIGR Internet server¹. Numbers are the gene identifiers in the HIBD database, where 234 means HI0234.#f: Indicate where two or more sequence fragments have been merged due to the identification of putative frameshifts. #m: Potentially incorrectly merged sequences (missing stop codon) are indicated by giving two functions separated by a '+'. 3D: Proteins for which significant sequence similarity was detected to a protein of known 3D structure allowing fairly accurate three-dimensional models to be built⁶. 18 (out of 148) of the new functions identified are the results of database matches with proteins that entered the databases between 16 May 1995 (the submission date of the TIGR analysis) and Aug 3, 1995. In addition we have classified an additional 55 new functions from database hits as 'tentative' and another 254 as 'marginal'. These results are accessible over the World Wide Web (<http://www.sander.embl-heidelberg.de/genequiz/haemophilus.html>) in the form of 'living' features which are automatically updated as new data arrives.

es, enhanced by frameshift detection, pattern searches, family analysis, structure prediction, knowledge navigation tools and information parsers.⁴

Proposed protein functions (see table and the TIGR table¹) are, for the most part, implied from the known functions of related proteins. Depending on the species, context and evolutionary distance, however, the similarity of function between related proteins varies substantially. The assigned functions should therefore be treated as plausible hypotheses (or "putative identification"¹), ranging from certain to approximate.

The 148 new functions include, among others, 65 enzymes, 16 transporters and 7 proteins involved in electron transfer. These have increased the overall assignment levels for *H. influenzae* to 9% - implied known 3D structure, 65% - clear functional assignment, 82% - clear database similarity hit. The distribution of functional roles, classified in nine classes similar to those of Riley⁵, is very similar to that of the currently available *Escherichia coli* sequences.

This analysis of *H. influenzae* protein sequences illustrates how automated computational sequence analysis can be rapid, exhaustive and accurate up to a level normally reached through extensive work of sequence analysis experts. Keeping pace with future advances in sequencing technology by imaginative informatics methods will be a continuing challenge.

**G. Casari,
M. A. Andrade,
P. Bork,[†]
J. Boyle,
A. Daruvar,
C. Ouzounis,[‡]
R. Schneider,
J. Tamames,[§]
A. Valencia,[§]
C. Sander,**

European Molecular Biology Laboratory
D-69012 Heidelberg, E.U.

Present Addresses

[†]Max-Delbrück Center for Molecular Medicine,
D-13112 Berlin, E.U.

[‡]SRI International, Menlo Park,
CA 94025-3493,
U.S.A.

[§]CNB-CSIC, Campus U. Autonoma,
Cantoblanco,
Madrid 28045, E.U.

Late experience alters vision

SIR — A visual target is said to 'pop out' if it is detected quickly and without serial search. Targets that differ from distractors by the presence of primitive visual features (for example colour, orientation) pop out whereas targets that differ based on other features do not¹. Treisman and Gelade¹ proposed that primitive features are processed in specialized modules, called feature maps. This hypothesis is consistent with neuroscientific evidence for spatially segregated cortical areas specialized for processing such features².

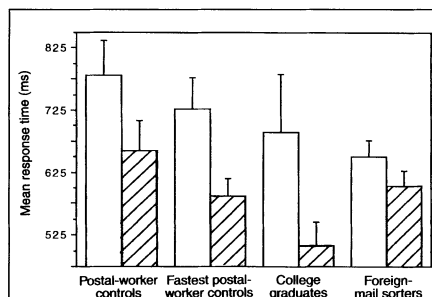
A similar pop-out effect occurs for letters and digits: a letter target is detected more efficiently among digits than among letters (and vice versa)³, suggesting that category information about alphanumeric stimuli is computed automatically and in parallel in much the same way as information about primitive visual features is processed. Consistent with this hypothesis, brain-damaged patients can be selectively impaired at letter recognition over number recognition⁴ and vice versa⁵, suggesting that letter and digit recognition are carried out by distinct maps. Recordings from human extrastriate visual cortex using chronically implanted electrodes also suggest segregation of letter and digit processing⁶.

Letter and digit recognition are not innate, so if the functional architecture of vision makes a distinction between letters and numbers, the environment must be helping to shape that architecture. How might this happen? The fact that neural learning is correlation-based suggests that correlations in the environment may

provide an answer. Stimuli within a category (letters) tend to occur together (in text) and are thus encoded in close temporal proximity, but stimuli between categories (letters and digits) co-occur less often. This statistical feature could interact with correlation-based learning in the brain to lead to maps for letter and digit recognition. We have implemented this hypothesis in a Hebbian neural network⁷ which, when presented with inputs that satisfy co-occurrence statistics, spontaneously self-organized to produce distinct letter and digit maps.

The co-occurrence hypothesis predicts a smaller category effect in subjects who regularly see letters and digits together. Such subjects should have less segregated maps and, as a result, letters should not pop out as much from digits. To test this prediction, we compared the effect in foreign-mail sorters at the Philadelphia Air Mail Facility with that in control subjects. Foreign-mail sorters spend roughly 4 hours per day processing Canadian zip codes in which letters and digits occur together (for example, M5S 1A4).

As predicted, foreign-mail sorters showed a reduced category effect compared with postal-worker controls, as measured both by the absolute difference in and ratio of response times in the letter-among-letters (LL) and letter-among-digits (LD) conditions (see figure). The sorters were, however, faster than controls, presumably because of their extensive experience with rapid tasks. To ensure that these results were not the result of a floor effect, we excluded the three slowest postal-worker controls (out of 16) for one analysis and used



Mean reaction times (with standard error bars) for correct responses on target-present trials in both letter-among-letter (LL; open bars) and letter-among-digit (LD; shaded bars) conditions for 16 postal-worker controls, the 13 fastest postal-worker controls, 8 college graduates and 10 foreign-mail sorters. All the foreign-mail sorters had been sorting foreign mail for at least 6 months before their participation. A computer presented subjects with a target letter followed by a circular array of letters or digits, and subjects were asked to press a key indicating whether or not the target was present. Details are based on experi-

ment 1 of ref. 13. *a*, Foreign-mail sorters compared with postal-worker controls. LD trials were reliably faster than LL trials ($t(25) = 4.367$, $P < 0.001$, one-tailed). Foreign-mail sorters showed a smaller alphanumeric category effect than did the postal-worker controls as measured both by the difference between the LL and LD conditions ($t(24) = 1.816$, $P < 0.05$, one-tailed) and by the LL/LD ratio ($t(24) = 1.864$, $P < 0.05$, one-tailed). *b*, Foreign-mail sorters against fastest postal-worker controls. LD trials were significantly faster than LL trials ($t(22) = 4.785$, $P < 0.001$, one-tailed) and the foreign-mail sorters showed a reduced category effect as measured by the LL-LD difference ($t(21) = 1.881$, $P < 0.05$, one-tailed) and by the LL/LD ratio ($t(21) = 2.060$, $P < 0.05$, one-tailed). *c*, Foreign-mail sorters against college graduates. Again, the LD condition was significantly faster than the LL condition ($t(17) = 3.208$, $P < 0.01$, one-tailed) and the foreign-mail sorters showed a reduced category effect as measured by the LL-LD difference ($t(16) = 2.271$, $P < 0.05$, one-tailed) and by the LL/LD ratio ($t(16) = 2.792$, $P < 0.01$, one-tailed).

1. Fleischmann, R.D. et al., *Science* **269** 496-512 (1995).

2. Scharf, M. et al. in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* (eds Altman, R., Brutlag, D., Karp, P., Lathrop, R. & Searls, D.) 348-353 (AAAI Press, Menlo Park, CA, 1994).

3. Bork, P. et al., *Molec. Microbiol.* **16** 955-967 (1995).

4. Etzold, T. & Argos P. *Comput. appl. Biosci.* **9** 49-57 (1993).

5. Riley, M., *Microbiol. Rev.* **57** 862 (1993).

6. Vriend, G., *J. molec. Graph.* **8** 52-56 (1990).