VOLUME 31 NUMBER 13 WEB SERVER ISSUE JULY 1, 2003

Nucleic Acids Research

NAR ONLINE
http://www.nar.oupjournals.org

OXFORD UNIVERSITY PRESS

ISSN 0305 1048 Coden NARHAD

Subscriptions

LIBRARY OF CONGRESS
JUL 7 2003
COPYRIGHT OFFICE
COPY

Nucleic Acids Research is published twice monthly (24 issues per year). Subscriptions are entered on a calendar year basis only and are available as a printed version (including online access) or as online access only at a discount of 10% (+VAT in UK). Prices include postage by surface mail or, for subscribers in the USA and Canada by air freight, or in India, Japan, Australia and New Zealand, by Air Speeded Post. Airmail rates are available on request.

Annual subscription rate (Volume 31, 2003):
Institutional
Print and Online site licence: UK and Europe £1365, Rest of World $2360.
Personal
Print and Online: UK and Europe £365, Rest of World $621.
Back volume prices are available on request. Please add sales tax to the prices quoted.

Orders. Orders and payments from, or on behalf of, subscribers in the various geographical areas shown below should be sent to the Press office indicated.

The Americas: Oxford University Press Inc., 2001 Evans Road, Cary, NC 27513, USA.

Rest of the World: Journals Subscriptions Department, Oxford University Press, Great Clarendon Street, Oxford OX2 6DP, UK.
Tel: (+441865 or 01865) 353907; Fax: (+441865) 353485;
Email: jnl.orders@oup.co.uk

Advertising. To advertise in *Nucleic Acids Research* contact Oxford University Press (US office) in the Americas or Oxford University Press (UK office) in the Rest of the World (see addresses above).

© Oxford University Press 2003. All rights reserved; no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise without either prior written permission of the Publishers, or a licence permitting restricted copying issued in the UK by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London WIP 9HE, or in the USA by the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, USA.

Back volumes of this journal are available in 16 mm microfilm, 35 mm microfilm and 105 microfiche from University Microfilms International, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, USA. Copies of articles published are also available from UMI.

Nucleic Acids Research (ISSN 0305-1048) is published twice monthly by Oxford University Press, Oxford, UK. Annual subscription price is US$2360. *Nucleic Acids Research* is distributed by Mercury International, 365 Blair Road, Avenel, NJ 07001, USA. Periodicals Postage Paid at Rahway, NJ, USA and additional entry points.

US POSTMASTER: send address corrections to *Nucleic Acids Research*, c/o Mercury Airfreight International Ltd, 365 Blair Road, Avenel, NJ 07001, USA.

Typeset and printed by Information Press Ltd, Oxford, UK on acid-free paper.

Cover: 3D-Jury model of the C-terminal RNA methyltransferase module of the large ORF1ab protein of the coronavirus isolated from patients suffering from SARS (Severe Acute Respiratory Syndrome). This kind of structure assignment is currently not feasible using standard sequence comparison methods.

Nucleic Acids Research

Contents

Volume 31 number 13, July 1, 2003

EDITORIAL		3289
Detection of reliable and unexpected protein fold predictions using 3D-Jury	K.Ginalski and L.Rychlewski	3291–3292
DSSPcont: continuous secondary structure assignments for proteins	P.Carter, C.A.F.Andersen and B.Rost	3293–3295
PROTINFO: secondary and tertiary protein structure prediction	L.-H.Hung and R.Samudrala	3296–3299
The PredictProtein server	B.Rost and J.Liu	3300–3304
GeneSilico protein structure prediction meta-server	M.A.Kurowski and J.M.Bujnicki	3305–3307
META-PP: single interface to crucial prediction servers	V.A.Eyrich and B.Rost	3308–3310
EVA: evaluation of protein structure prediction servers	I.Y.Y.Koh, V.A.Eyrich, M.A.Marti-Renom, D.Przybylski, M.S.Madhusudhan, N.Eswar, O.Graña, F.Pazos, A.Valencia, A.Sali and B.Rost	3311–3315
VADAR: a web server for quantitative evaluation of protein structure quality	L.Willard, A.Ranjan, H.Zhang, H.Monzavi, R.F.Boyko, B.D.Sykes and D.S.Wishart	3316–3319
ESPript/ENDscript: extracting and rendering sequence and 3D information from atomic structures of proteins	P.Gouet, X.Robert and E.Courcelle	3320–3323
WebFEATURE: an interactive web tool for identifying and visualizing functional sites on macromolecular structures	M.P.Liang, D.R.Banatao, T.E.Klein, D.L.Brutlag and R.B.Altman	3324–3327
3MATRIX and 3MOTIF: a protein structure visualization system for conserved sequence motifs	S.P.Bennett, L.Lu and D.L.Brutlag	3328–3332
Motif3D: relating protein sequence motifs to 3D structure	A.Gaulton and T.K.Attwood	3333–3336
LOC3D: annotate sub-cellular localization for protein structures	R.Nair and B.Rost	3337–3340
Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures	A.Stark and R.B.Russell	3341–3344
NCI: a server to identify non-canonical interactions in protein structures	M.M.Babu	3345–3348
MolSurfer: a macromolecular interface navigator	R.R.Gabdoulline, R.C.Wade and D.Walther	3349–3351
CASTp: Computed Atlas of Surface Topography of proteins	T.A.Binkowski, S.Naghibzadeh and J.Liang	3352–3355

Continued

Contents (Continued)

Volume 31 number 13, July 1, 2003

	SEM (Symmetry Equivalent Molecules): a web-based GUI to generate and visualize the macromolecules	A.S.Z.Hussain, Ch.K.Kumar, C.K.Rajesh, S.S.Sheik and K.Sekar	3356–3358
	Servers for sequence–structure relationship analysis and prediction	Z.Dosztányi, C.Magyar, G.E.Tusnády, M.Cserző, A.Fiser and I.Simon	3359–3363
	POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level	L.Cavallo, J.Kleinjung and F.Fraternali	3364–3366
	MATRAS: a program for protein 3D structure comparison	T.Kawabata	3367–3369
	LGA: a method for finding 3D similarities in protein structures	A.Zemla	3370–3374
	Tools for comparative protein structure modeling and analysis	N.Eswar, B.John, N.Mirkovic, A.Fiser, V.A.Ilyin, U.Pieper, A.C.Stuart, M.A.Marti-Renom, M.S.Madhusudhan, B.Yerkovich and A.Sali	3375–3380
	SWISS-MODEL: an automated protein homology-modeling server	T.Schwede, J.Kopp, N.Guex and M.C.Peitsch	3381–3385
	STING Millennium: a web-based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence	G.Neshich, R.C.Togawa, A.L.Mancini, P.R.Kuser, M.E.B.Yamagishi, G.Pappas Jr, W.V.Torres, T.F.eCampos, L.L.Ferreira, F.M.Luna, A.G.Oliveira, R.T.Miura, M.K.Inoue, L.G.Horita, D.F.de Souza, F.Dominiquini, A.Álvaro, C.S.Lima, F.O.Ogawa, G.B.Gomes, J.F.Palandrani, G.F.dos Santos, E.M.de Freitas, A.R.Mattiuz, I.C.Costa, C.L.de Almeida, S.Souza, C.Baudet and R.H.Higa	3386–3392
	Integrated databanks access and sequence/structure analysis services at the PBIL	G.Perrière, C.Combet, S.Penel, C.Blanchet, J.Thioulouse, C.Geourjon, J.Grassot, C.Charavay, M.Gouy, L.Duret and G.Deléage	3393–3399
	NRSAS: Nuclear Receptor Structure Analysis Servers	E.Bettler, R.Krause, F.Horn and G.Vriend	3400–3403
	SSEP: secondary structural elements of proteins	V.Shanthi, P.Selvarani, Ch.K.Kumar, C.S.Mohire and K.Sekar	3404–3405
S	**Mfold web server for nucleic acid folding and hybridization prediction**	M.Zuker	3406–3415
	RNAsoft: a suite of RNA secondary structure prediction and design software tools	M.Andronescu, R.Aguirre-Hernández, A.Condon and H.H.Hoos	3416–3422
S	**Pfold: RNA secondary structure prediction using stochastic context-free grammars**	B.Knudsen and J.Hein	3423–3428
	Vienna RNA secondary structure server	I.L.Hofacker	3429–3431
	PseudoViewer2: visualization of RNA pseudoknots of any type	K.Han and Y.Byun	3432–3440
	A software tool-box for analysis of regulatory RNA elements	P.Bengert and T.Dandekar	3441–3445
	GPRM: a genetic programming approach to finding common RNA secondary structure elements	Y.-J.Hu	3446–3449

Continued

Contents (Continued)

Volume 31 number 13, July 1, 2003

Tools for the automatic identification and classification of RNA base pairs	H.Yang, F.Jossinet, N.Leontis, L.Chen, J.Westbrook, H.Berman and E.Westhof	3450–3460
GEPAS: a web-based resource for microarray gene expression data analysis	J.Herrero, F.Al-Shahrour, R.Díaz-Uriarte, A.Mateos, J.M.Vaquerizas, J.Santoyo and J.Dopazo	3461–3467
INCLUSive: a web portal and service registry for microarray and regulatory sequence analysis	B.Coessens, G.Thijs, S.Aerts, K.Marchal, F.De Smet, K.Engelen, P.Glenisson, Y.Moreau, J.Mathys and B.De Moor	3468–3470
GenePublisher: automated analysis of DNA microarray data	S.Knudsen, C.Workman, T.Sicheritz-Ponten and C.Friis	3471–3476
ExpressYourself: a modular platform for processing and visualizing microarray data	N.M.Luscombe, T.E.Royce, P.Bertone, N.Echols, C.E.Horak, J.T.Chang, M.Snyder and M.Gerstein	3477–3482
ChipInfo: software for extracting gene annotation and gene ontology information for microarray analysis	S.Zhong, C.Li and W.H.Wong	3483–3486
REDUCE: an online tool for inferring *cis*-regulatory elements and transcriptional module activities from microarray data	C.Roven and H.J.Bussemaker	3487–3490
Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays	H.B.Nielsen, R.Wernersson and S.Knudsen	3491–3496
Multiple sequence alignment with the Clustal series of programs	R.Chenna, H.Sugawara, T.Koike, R.Lopez, T.J.Gibson, D.G.Higgins and J.D.Thompson	3497–3500
CLOURE: Clustal Output Reformatter, a program for reformatting ClustalX/ClustalW outputs for SNP analysis and molecular systematics	D.K.Kohli and A.K.Bachhawat	3501–3502
Tcoffee@igs: a web server for computing, evaluating and combining multiple sequence alignments	O.Poirot, E.O'Toole and C.Notredame	3503–3506
SLAM web server for comparative gene finding and alignment	S.Cawley, L.Pachter and M.Alexandersson	3507–3509
Theatre: a software tool for detailed comparative analysis and visualization of genomic sequence	Y.J.K.Edwards, T.J.Carver, T.Vavouri, M.Frith, M.J.Bishop and G.Elgar	3510–3517
MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences	S.Schwartz, L.Elnitski, M.Li, M.Weirauch, C.Riemer, A.Smit, NISC Comparative Sequencing Program, E.D.Green, R.C.Hardison and W.Miller	3518–3524
MAVID multiple alignment server	N.Bray and L.Pachter	3525–3526
EnteriX 2003: visualization tools for genome alignments of Enterobacteriaceae	L.Florea, M.McClelland, C.Riemer, S.Schwartz and W.Miller	3527–3532
MGAlignIt: a web service for the alignment of mRNA/EST and genomic sequences	B.T.K.Lee, T.W.Tan and S.Ranganathan	3533–3536
RevTrans: multiple alignment of coding DNA from aligned amino acid sequences	R.Wernersson and A.G.Pedersen	3537–3539
S **PromH: promoters identification using orthologous genomic sequences**	V.V.Solovyev and I.A.Shahmuradov	3540–3545

Continued

Contents (Continued)

Volume 31 number 13, July 1, 2003

S	**FIE2: a program for the extraction of genomic DNA sequences around the start and translation initiation site of human genes**	A.Chong, G.Zhang and V.B.Bajic	3546–3553
	PromoSer: a large-scale mammalian promoter and transcription start site identification service	A.S.Halees, D.Leyfer and Z.Weng	3554–3559
S	**Dragon Gene Start Finder identifies approximate locations of the 5′ ends of genes**	V.B.Bajic and S.H.Seah	3560–3563
	ETOPE: evolutionary test of predicted exons	A.Nekrutenko, W.-Y.Chung and W.H.Li	3564–3567
	ESEfinder: a web resource to identify exonic splicing enhancers	L.Cartegni, J.Wang, Z.Zhu, M.Q.Zhang and A.R.Krainer	3568–3571
	SiteSeer: visualisation and analysis of transcription factor binding sites in nucleotide sequences	P.E.Boardman, S.G.Oliver and S.J.Hubbard	3572–3575
	MATCH™: a tool for searching transcription factor binding sites in DNA sequences	A.Kel, E.Gößling, I.Reuter, E.Cheremushkin, O.Kel-Margoulis and E.Wingender	3576–3579
	Gibbs Recursive Sampler: finding transcription factor binding sites	W.Thompson, E.C.Rouchka and C.E.Lawrence	3580–3585
	YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation	S.Sinha and M.Tompa	3586–3588
	Target Explorer: an automated tool for the identification of new target genes for a specified set of transcription factors	A.Sosinsky, C.P.Bonin, R.S.Mann and B.Honig	3589–3592
	Regulatory Sequence Analysis Tools	J.van Helden	3593–3596
	GeneSeqer@PlantGDB: gene structure prediction in plant genomes	S.D.Schlueter, Q.Dong and V.Brendel	3597–3600
	GlimmerM, Exonomy and Unveil: three *ab initio* eukaryotic genefinders	W.H.Majoros, M.Pertea, C.Antonescu and S.L.Salzberg	3601–3604
S	**Dragon ERE Finder version 2: a tool for accurate detection and analysis of estrogen response elements in vertebrate genomes**	V.B.Bajic, S.L.Tan, A.Chong, S.Tang, A.Ström, J.-A.Gustafsson, C.-Y.Lin and E.T.Liu	3605–3607
	PatSearch: a program for the detection of patterns and structural motifs in nucleotide sequences	G.Grillo, F.Licciulli, S.Liuni, E.Sbisà and G.Pesole	3608–3612
	PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria	J.L.Gardy, C.Spencer, K.Wang, M.Ester, G.E.Tusnády, I.Simon, S.Hua, K.deFays, C.Lambert, K.Nakai and F.S.L.Brinkman	3613–3617
	Signal search analysis server	G.Ambrosini, V.Praz, V.Jagannathan and P.Bucher	3618–3620
	MHCPred: a server for quantitative prediction of peptide–MHC binding	P.Guan, I.A.Doytchinova, C.Zygouri and D.R.Flower	3621–3624

Continued

Contents (Continued)

	ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins	P.Puntervoll, R.Linding, C.Gemünd, S.Chabanis-Davidson, M.Mattingsdal, S.Cameron, D.M.A.Martin, G.Ausiello, B.Brannetti, A.Costantini, F.Ferrè, V.Maselli, A.Via, G.Cesareni, F.Diella, G.Superti-Furga, L.Wyrwicz, C.Ramu, C.McGuigan, R.Gudavalli, I.Letunic, P.Bork, L.Rychlewski, B.Küster, M.Helmer-Citterich, W.N.Hunter, R.Aasland and T.J.Gibson	3625–3630
	Prediction of lipid posttranslational modifications and localization signals from protein sequences: big-Π, NMT and PTS1	F.Eisenhaber, B.Eisenhaber, W.Kubina, S.Maurer-Stroh, G.Neuberger, G.Schneider and M.Wildpaner	3631–3634
	Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs	J.C.Obenauer, L.C.Cantley and M.B.Yaffe	3635–3641
	Static benchmarking of membrane helix predictions	A.Kernytsky and B.Rost	3642–3644
	The web server of IBM's Bioinformatics and Pattern Discovery group	T.Huynh, I.Rigoutsos, L.Parida, D.Platt and T.Shibuya	3645–3650
	Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN	D.Farré, R.Roset, M.Huerta, J.E.Adsuara, L.Roselló, M.M.Albà and X.Messeguer	3651–3653
	SEARCHPKS: a program for detection and analysis of polyketide synthase domains	G.Yadav, R.S.Gokhale and D.Mohanty	3654–3658
S	**MAK, a computational tool kit for automated MITE analysis**	G.Yang and T.C.Hall	3659–3665
	Cluster-Buster: finding dense clusters of motifs in DNA sequences	M.C.Frith, M.C.Li and Z.Weng	3666–3668
	SIC: a tool to detect short inverted segments in a biological sequence	D.Robelin, H.Richard and B.Prum	3669–3671
	`mreps`: efficient and flexible detection of tandem repeats in DNA	R.Kolpakov, G.Bana and G.Kucherov	3672–3678
	SPA: simple web tool to assess statistical significance of DNA patterns	H.Richard and G.Nuel	3679–3681
	TRACTS: a program to map oligopurine.oligopyrimidine and other binary DNA tracts	M.Gal, T.Katz, A.Ovadia and G.Yagil	3682–3685
	DNA analysis servers: plot.it, bend.it, model.it and IS	K.Vlahoviček, L.Kaján and S.Pongor	3686–3687
	NEBcutter: a program to cleave DNA with restriction enzymes	T.Vincze, J.Posfai and R.J.Roberts	3688–3691
	SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence	C.Z.Cai, L.Y.Han, Z.L.Ji, X.Chen and Y.Z.Chen	3692–3697
	BPROMPT: a consensus server for membrane protein prediction	P.D.Taylor, T.K.Attwood and D.R.Flower	3698–3700
	GlobPlot: exploring protein sequences for globularity and disorder	R.Linding, R.B.Russell, V.Neduva and T.J.Gibson	3701–3708

Continued

Contents (Continued)

	iSPOT: a web tool to infer the interaction specificity of families of protein modules	B.Brannetti and M.Helmer-Citterich	3709–3711
	Automated Gene Ontology annotation for anonymous sequence data	S.Hennig, D.Groth and H.Lehrach	3712–3715
	ESTAnnotator: a tool for high throughput EST annotation	A.Hotz-Wagenblatt, T.Hankeln, P.Ernst, K.-H.Glatting, E.R.Schmidt and S.Suhai	3716–3719
	Phydbac (phylogenomic display of bacterial genes): an interactive resource for the annotation of bacterial genomes	F.Enault, K.Suhre, O.Poirot, C.Abergel and J.-M.Claverie	3720–3722
	AMIGene: Annotation of MIcrobial Genes	S.Bocs, S.Cruveiller, D.Vallenet, G.Nuel and C.Médigue	3723–3726
	AHMII: Agent to Help Microbial Information Integration	H.Sugawara and S.Miyazaki	3727–3728
S	**DNannotator: annotation software tool kit for regional genomic sequences**	C.Liu, T.I.Bonner, T.Nguyen, J.L.Lyons, S.L.Christian and E.S.Gershon	3729–3735
	Bioverse: functional, structural and contextual annotation of proteins and proteomes	J.McDermott and R.Samudrala	3736–3737
S	**FrameD: a flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences**	T.Schiex, J.Gouzy, A.Moisan and Y.de Oliveira	3738–3741
S	**EuGène'Hom: a generic similarity-based gene finder using multiple homologous sequences**	S.Foissac, P.Bardou, A.Moisan, M.-J.Cros and T.Schiex	3742–3745
	PROBEmer: a web-based software tool for selecting optimal DNA oligos	S.J.Emrich, M.Lowe and A.L.Delcher	3746–3750
	Primer Design Assistant (PDA): a web-based primer design tool	S.H.Chen, C.Y.Lin, C.S.Cho, C.Z.Lo and C.A.Hsiung	3751–3754
	DePIE: Designing Primers for Protein Interaction Experiments	G.Lu, M.Hallett, S.Pollock and D.Thomas	3755–3757
	OligoDesign: optimal design of LNA (locked nucleic acid) oligonucleotide capture probes for gene expression profiling	N.Tolstrup, P.S.Nielsen, J.G.Kolberg, A.M.Frankel, H.Vissing and S.Kauppinen	3758–3762
	CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design	T.M.Rose, J.G.Henikoff and S.Henikoff	3763–3766
S	**RNA-related tools on the Bielefeld Bioinformatics Server**	A.Sczyrba, J.Krüger, H.Mersch, S.Kurtz and R.Giegerich	3767–3770
	SIRW: a web server for the Simple Indexing and Retrieval System that combines sequence motif searches with keyword searches	C.Ramu	3771–3774
	Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate	S.Draghici, P.Khatri, P.Bhavsar, A.Shah, S.A.Krawetz and M.A.Tainsky	3775–3781
	Swiss EMBnet node web server	L.Falquet, L.Bordoli, V.Ioannidis, M.Pagni and C.V.Jongeneel	3782–3783

Continued

Contents (Continued)

Volume 31 number 13, July 1, 2003

	ExPASy: the proteomics server for in-depth protein knowledge and analysis	E.Gasteiger, A.Gattiker, C.Hoogland, I.Ivanyi, R.D.Appel and A.Bairoch	3784–3788
	UniqueProt: creating representative protein sequence sets	S.Mika and B.Rost	3789–3791
	BLAST2SRS, a web server for flexible retrieval of related protein sequences in the SWISS-PROT and SPTrEMBL databases	K.Bimpikis, A.Budd, R.Linding and T.J.Gibson	3792–3794
	WU-Blast2 server at the European Bioinformatics Institute	R.Lopez, V.Silventoinen, S.Robinson, A.Kibria and W.Gish	3795–3798
	OntoBlast function: from sequence similarities directly to potential functional annotations by ontology terms	G.Zehetner	3799–3803
	ORFeus: detection of distant homology using sequence profiles and predicted secondary structure	K.Ginalski, J.Pas, L.S.Wyrwicz, M.von Grotthuss, J.M.Bujnicki and L.Rychlewski	3804–3807
	PARSESNP: a tool for the analysis of nucleotide polymorphisms	N.E.Taylor and E.A.Greene	3808–3811
	SIFT: predicting amino acid changes that affect protein function	P.C.Ng and S.Henikoff	3812–3814
S	**ActionMap: a web-based software that automates loci assignments to framework maps**	G.Albini, M.Falque and J.Joets	3815–3818
	WEB-THERMODYN: sequence analysis software for profiling DNA helical stability	Y.Huang and D.Kowalski	3819–3821
	NEWT, a new taxonomy portal	I.Q.H.Phan, S.F.Pilbout, W.Fleischmann and A.Bairoch	3822–3823
	Comprehensive quantitative analyses of the effects of promoter sequence elements on mRNA transcription	M.Lapidot and Y.Pilpel	3824–3828
	PipeAlign: a new toolkit for protein family analysis	F.Plewniak, L.Bianchetti, Y.Brelivet, A.Carles, F.Chalmel, O.Lecompte, T.Mochel, L.Moulinier, A.Muller, J.Muller, V.Prigent, R.Ripp, J.-C.Thierry, J.D.Thompson, N.Wicker and O.Poch	3829–3832
	NORSp: predictions of long regions without regular secondary structure	J.Liu and B.Rost	3833–3835
	Biological SOAP servers and web services provided by the public sequence data bank	H.Sugawara and S.Miyazaki	3836–3839
	FootPrinter: a program designed for phylogenetic footprinting	M.Blanchette and M.Tompa	3840–3842
	GeneFizz: a web tool to compare genetic (coding/non-coding) and physical (helix/coil) segmentations of DNA sequences. Gene discovery and evolutionary perspectives	E.Yeramian and L.Jones	3843–3849
S	**Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes**	N.Beerenwinkel, M.Däumer, M.Oette, K.Korn, D.Hoffmann, R.Kaiser, T.Lengauer, J.Selbig and H.Walter	3850–3855

Continued

Contents (Continued)

Volume 31 number 13, July 1, 2003

	Building protein diagrams on the web with the residue-based diagram editor RbDe	L.Skrabanek, F.Campagne and H.Weinstein	3856–3858
	CRP: Cleavage of Radiolabeled Phosphoproteins	A.J.Mackey, T.A.J.Haystead and W.R.Pearson	3859–3861
S	**JVirGel: calculation of virtual two-dimensional protein gels**	K.Hiller, M.Schobert, C.Hundertmark, D.Jahn and R.Münch	3862–3865
	Update on XplorMed: a web server for exploring scientific literature	C.Perez-Iratxeta, A.J.Pérez, P.Bork and M.A.Andrade	3866–3868

AUTHOR INDEX

S, Supplementary Material available at NAR Online

This material may be protected by Copyright law (Title 17 U.S. Code)

3812–3814 Nucleic Acids Research, 2003, Vol. 31, No. 13
DOI: 10.1093/nar/gkg509

SIFT: predicting amino acid changes that affect protein function

Pauline C. Ng and Steven Henikoff*

Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N A1-162, Seattle, WA 98109, USA

Received January 4, 2003; Revised and Accepted February 28, 2003

ABSTRACT

Single nucleotide polymorphism (SNP) studies and random mutagenesis projects identify amino acid substitutions in protein-coding regions. Each substitution has the potential to affect protein function. SIFT (Sorting Intolerant From Tolerant) is a program that predicts whether an amino acid substitution affects protein function so that users can prioritize substitutions for further study. We have shown that SIFT can distinguish between functionally neutral and deleterious amino acid changes in mutagenesis studies and on human polymorphisms. SIFT is available at http://blocks.fhcrc.org/sift/SIFT.html.

INTRODUCTION

Single nucleotide polymorphisms (SNPs) are used as markers in linkage and association studies to detect which regions in the human genome may be involved in disease. SNPs in coding and regulatory regions may be implicated in disease themselves. Non-synonymous SNPs that lead to an amino acid change in the protein product are of major interest, because amino acid substitutions currently account for approximately half of the known gene lesions responsible for human inherited disease (1). SIFT (Sorting Intolerant From Tolerant) uses sequence homology to predict whether an amino acid substitution will affect protein function and hence, potentially alter phenotype (2,3).

SIFT has been applied to human variant databases and was able to distinguish mutations involved in disease from neutral polymorphisms (3). Assuming that disease-causing amino acid substitutions are damaging to protein function, we applied SIFT to a database of missense substitutions associated with or involved in disease (4). SIFT predicted 69% to be damaging. When SIFT was applied to the non-synonymous SNPs in dbSNP (5), a database of putative SNPs, 25% of the variants were predicted to be deleterious. This was similar to SIFT's 20% false positive error which suggested that most non-synonymous SNPs are functionally neutral. Furthermore, a subset of the variants from dbSNP predicted to affect function were involved in disease which confirmed SIFT sensitivity.

The SIFT algorithm relies solely on sequence for prediction, yet performs similarly to tools that use structure (3,6–8). An advantage of not requiring structure is that a larger number of substitutions can be predicted on. Of the non-synonymous SNPs identified by the SNP Consortium, 74% were sufficiently similar to homologs in protein sequence databases for SIFT prediction. The number of substitutions that SIFT can predict on is expected to increase as more genomes are sequenced and more protein sequences become available.

SIFT PREDICTION METHOD

SIFT presumes that important amino acids will be conserved in the protein family, and so changes at well-conserved positions tend to be predicted as deleterious. For example, if a position in an alignment of a protein family only contains the amino acid isoleucine, it is presumed that substitution to any other amino acid is selected against and that isoleucine is necessary for protein function. Therefore, a change to any other amino acid will be predicted to be deleterious to protein function. If a position in an alignment contains the hydrophobic amino acids isoleucine, valine and leucine, then SIFT assumes, in effect, that this position can only contain amino acids with hydrophobic character. At this position, changes to other hydrophobic amino acids are usually predicted to be tolerated but changes to other residues (such as charged or polar) will be predicted to affect protein function.

To predict whether an amino acid substitution in a protein will affect protein function, SIFT considers the position at which the change occurred and the type of amino acid change. Given a protein sequence, SIFT chooses related proteins and obtains an alignment of these proteins with the query. Based on the amino acids appearing at each position in the alignment, SIFT calculates the probability that an amino acid at a position is tolerated conditional on the most frequent amino acid being tolerated. If this normalized value is less than a cutoff, the substitution is predicted to be deleterious (2). The SIFT algorithm and software have been described previously (2,3).

SIFT WEBSITE

Input

Users can obtain predictions for amino acid changes of interest at http://www.blocks.fhcrc.org/sift/SIFT.html. From this page,

*To whom correspondence should be addressed. Tel: +1 2066674515; Fax: +1 2066675889; Email: steveh@fhcrc.org

Nucleic Acids Research, Vol. 31, No. 13 © Oxford University Press 2003; all rights reserved

```
Substitution at pos 1426 from S to P is predicted to AFFECT PROTEIN FUNCTION with a score of 0.02.
    Median sequence conservation: 2.90
    Sequences represented at this position:26

Substitution at pos 1432 from N to K is predicted to be TOLERATED with a score of 0.08.
    Median sequence conservation: 2.90
    Sequences represented at this position:26

Substitution at pos 1445 from D to N is predicted to AFFECT PROTEIN FUNCTION with a score of 0.01.
    Median sequence conservation: 3.66
    Sequences represented at this position:21
WARNING!!  This substitution may have been predicted to affect function just because
 the sequences used were not diverse enough.  There is LOW CONFIDENCE in this prediction.
```

Figure 1. An example of SIFT prediction on amino acid changes in a protein. Substitutions with score less than 0.05 are predicted to affect protein function. In the last prediction, the median conservation of the sequences does not meet the threshold so a warning is issued.

there are links to three submission pages which allow users different levels of involvement in order to control the quality of their predictions.

For minimal involvement, users can simply submit their protein sequences and amino acid substitutions. In its fully automated mode, SIFT will search for protein sequences homologous to the query protein and based on these sequences, calculate probabilities for each possible amino acid change. Users can select from among SWISS-PROT, SWISS-PROT/TrEMBL, or NCBI's non-redundant protein databases for SIFT to search (4,9).

Although SIFT can choose sequences automatically, better prediction results may be obtained when all of the sequences that are provided are orthologous to the query protein. This is because inclusion of paralogous sequences confounds prediction at residues conserved only among the orthologues. If a user already has sequences that are thought to be functionally similar to the protein of interest, these sequences can be directly submitted and SIFT's step for choosing sequences skipped. Given the query protein and homologous sequences, SIFT obtains the alignment.

If regions are misaligned, SIFT will not recognize conserved positions and therefore miss potentially damaging substitutions. For best prediction quality, a third mode of operation allows users to submit their own alignments.

Output

Predictions are given for all 20 possible amino acid changes at each position in the protein. The alignment is also returned so that users can examine the sequences used for prediction and modify them for resubmission. This option is also useful for removing uncertain, erroneous and misaligned sequences from alignment output generated by SIFT in its automatic mode.

For amino acid substitutions submitted by the user, a more detailed synopsis is provided (Fig. 1). The score is the normalized probability that the amino acid change is tolerated. SIFT predicts substitutions with scores less than 0.05 as deleterious. Some SIFT users have found that substitutions with scores less than 0.1 provide better sensitivity for detecting deleterious SNPs (Cornelia Ulrich, personal communication and 10). The quantitative score allows users to prioritize their amino acid changes by ranking them from the lowest scores to the highest.

Confidence in a substitution predicted to be deleterious depends on the diversity of the sequences in the alignment. If the sequences used for prediction are closely related, then many positions will appear conserved and SIFT will predict most substitutions to affect protein function. This leads to a high false positive error where functionally neutral substitutions are predicted to be deleterious.

To alert the user to these situations, SIFT calculates the median conservation value which measures the diversity of the sequences in the alignment. Conservation, as measured by information content (11), is calculated for each position in the alignment and the median of these values is obtained. Conservation ranges from $\log_2 20$ ($=4.32$), when a position is completely conserved and only one amino acid is observed, to zero, when all 20 amino acids are observed at a position. By default, SIFT builds alignments with a median conservation value of 3.0. Predictions based on sequence alignments with higher median conservation values are less diverse and will have a higher false positive error (Fig. 2).

Even if there are few homologous sequences available, SIFT performs better than simply predicting non-conservative amino acid substitutions as deleterious, where non-conservative changes are defined as having negative scores in an amino acid substitution scoring matrix. We have shown that with only one sequence homologous to the test protein, SIFT can predict twice as many neutral substitutions correctly compared to a substitution scoring matrix (2). Even with few homologous sequences, there will be positions that differ between the test protein and the other sequences. Depending on the amino acids appearing at these positions, SIFT may predict these positions to be unimportant for protein function. This additional information can eliminate functionally neutral substitutions and increase selectivity to deleterious substitutions.

In summary, a large number of substitutions can be obtained from mutagenesis projects, SNP datasets, and changes between closely related organisms. When it is not feasible to conduct experiments on all substitutions, SIFT and other similar prediction tools (13) may be useful in prioritizing which

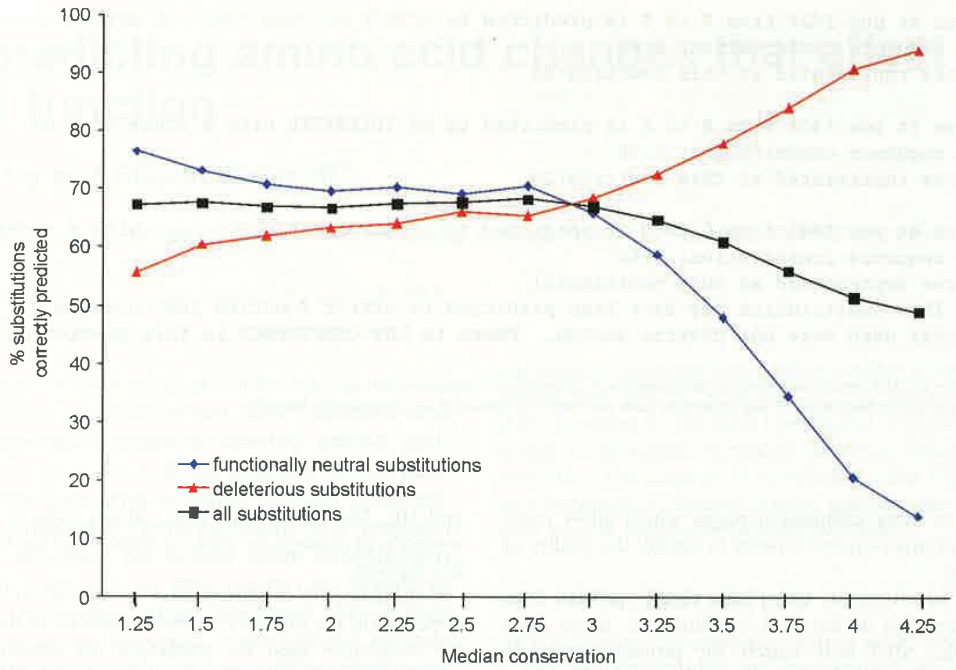

Figure 2. Prediction depends on the diversity of the sequences used in the alignment. Percentage of substitutions correctly predicted is based on over 4000 substitutions that were assayed throughout the LacI protein of *Escherichia coli* (2,12). When the sequences in the alignment used for prediction are closely related (high median conservation) then many positions appear conserved and important for function. In this situation, prediction accuracy on deleterious substitutions is high but many functionally neutral substitutions are erroneously predicted to be deleterious. To obtain an alignment with a specified median conservation, the LacI protein sequence of *E.coli* was submitted to the SIFT website and the median conservation setting adjusted. Because the homologous sequences available are distantly related to *E.coli* LacI, alignments with higher median conservation values could not be obtained. In order to obtain alignments with median conservation values more than 3.25, closely related sequences were simulated by starting with an alignment of identical *E.coli* LacI sequences. A position and a sequence were randomly selected from the LacI alignment with median conservation 2.75. The amino acid corresponding to this location was substituted in the starting alignment. Amino acids continued to be randomly selected and substituted until the desired median conservation was met. The simulated alignment was then evaluated for its performance as previously described (2) and the plotted value is the average performance of 100 simulated alignments.

changes affect protein function and may contribute to phenotypic differences.

ACKNOWLEDGEMENTS

We thank Jorja Henikoff for advice and encouragement. This work was supported by a grant from NIH (GM29009).

REFERENCES

1. Krawczak,M., Ball,E.V., Fenton,I., Stenson,P.D., Abeysinghe,S., Thomas,N. and Cooper,D.N. (2000) Human gene mutation database-a biomedical information and research resource. *Hum. Mutat.*, **15**, 45–51.
2. Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.
3. Ng,P.C. and Henikoff,S. (2002) Accounting for human polymorphisms predicted to affect protein function. *Genome Res.*, **12**, 436–446.
4. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
5. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
6. Sunyaev,S., Ramensky,V., Koch,I., Lathe,W.,III, Kondrashov,A.S. and Bork,P. (2001) Prediction of deleterious human alleles. *Hum. Mol. Genet.*, **10**, 591–597.
7. Chasman,D. and Adams,R.M. (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.*, **307**, 683–706.
8. Saunders,C.T. and Baker,D. (2002) Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol.*, **322**, 891–901.
9. Wheeler,D.L., Church,D.M., Lash,A.E., Leipe,D.D., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Tatusova,T.A., Wagner,L. and Rapp,B.A. (2002) Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res.*, **30**, 13–16.
10. Leabman,M.K., Huang,C.C., DeYoung,J., Carlson,E.J., Taylor,T., de la Cruz,M., Johns,S.J., Stryke,D., Kawamoto,M., Urban,T.J., *et al.* (2003) Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. *Proc. Natl Acad. Sci. USA*, in press.
11. Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.
12. Pace,H.C., Kercher,M.A., Lu,P., Markiewicz,P., Miller,J.H., Chang,G. and Lewis,M. (1997) Lac repressor genetic map in real space. *Trends Biochem. Sci.*, **22**, 334–339.
13. Ramensky,V., Bork,P. and Sunyaev, S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.