Protein Science (1995), *4*:1618-1632. Cambridge University Press. Printed in the USA.

Gibbs motif sampling: Detection of bacterial outer membrane protein repeats

ANDREW F. NEUWALD,[1] JUN S. LIU,[2] AND CHARLES E. LAWRENCE[1,3]

[1] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894
[2] Department of Statistics, Stanford University, Stanford, California 94305
[3] Biometrics Laboratory, Wadsworth Center for Laboratories and Research, New York State Department of Health, Albany, New York 12201

(RECEIVED April 4, 1995; ACCEPTED May 24, 1995)

Abstract

The detection and alignment of locally conserved regions (motifs) in multiple sequences can provide insight into protein structure, function, and evolution. A new Gibbs sampling algorithm is described that detects motif-encoding regions in sequences and optimally partitions them into distinct motif models; this is illustrated using a set of immunoglobulin fold proteins. When applied to sequences sharing a single motif, the sampler can be used to classify motif regions into related submodels, as is illustrated using helix-turn-helix DNA-binding proteins. Other statistically based procedures are described for searching a database for sequences matching motifs found by the sampler. When applied to a set of 32 very distantly related bacterial integral outer membrane proteins, the sampler revealed that they share a subtle, repetitive motif. Although BLAST (Altschul SF et al., 1990, *J Mol Biol 215*:403-410) fails to detect significant pairwise similarity between any of the sequences, the repeats present in these outer membrane proteins, taken as a whole, are highly significant (based on a generally applicable statistical test for motifs described here). Analysis of bacterial porins with known trimeric β-barrel structure and related proteins reveals a similar repetitive motif corresponding to alternating membrane-spanning β-strands. These β-strands occur on the membrane interface (as opposed to the trimeric interface) of the β-barrel. The broad conservation and structural location of these repeats suggests that they play important functional roles.

Keywords: Bayesian inference; multiple alignment algorithms; outer membrane proteins; pattern recognition; porins; protein motifs; statistical significance; Wilcoxon signed rank test

Sequence similarity, found using either pairwise alignment, multiple alignment, or motif detection methods, often yields the first clues to protein structure and function. The detection of weakly conserved patterns (motifs) among distantly related sequences can be particularly informative because they often correspond to structurally or functionally important residues. Such information is useful for targeting specific sites for in vitro mutagenesis and in classifying diverse proteins according to implied structural and/or mechanistic similarities. Alignment profiles of conserved regions have been useful for detecting very distant relationships (Gribskov et al., 1987, 1990; Luthy et al., 1994) and, when compiled into profile databases, can be useful for screening new sequences for motifs (Henikoff & Henikoff, 1991, 1994).

Pairwise and multiple sequence analysis methods for detecting similarity between relatively closely related sequences have been available for some time now (Needleman & Wunsch, 1970; Smith & Waterman, 1981; Pearson & Lipman, 1988; Altschul et al., 1990; for a review of multiple alignment methods see Chan et al., 1992). Only more recently, however, have efficient methods been developed that can detect subtle similarities common to large sets of distantly related or (possibly) evolutionarily unrelated sequences (Lawrence et al., 1993; Neuwald & Green, 1994). The development of these methods has been motivated by the current rapid increase in sequence data because relatively large sets (containing, for example, more than 15 sequences) are needed for weakly conserved patterns to reach statistical significance.

Lawrence et al. (1993) describe a Gibbs sampling strategy for detecting conserved patterns in multiple sequences that is a stochastic analog of earlier expectation–maximization methods (Lawrence & Reilly, 1990; Cardon & Stormo, 1992) and that is closely related to (EM-based) hidden Markov model multiple sequence alignment methods (Baldi et al., 1994; Krogh et al.,

Reprint requests to: Andrew F. Neuwald, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894; e-mail: neuwald@ncbi.nlm.nih.gov.

Abbreviations: EM, expectation–maximization; HMM, hidden Markov model; hth, helix-turn-helix; iomp, integral outer membrane protein; ipp, information per parameter; MSP, maximal segment pair.

1994), which, unlike the Gibbs sampler, permit gaps anywhere in the sequences. This Gibbs sampler (which is referred to here as the site sampler) addresses the problem of finding motifs when the number of occurrences of each motif in each sequence is assumed. Using such prior information, when justified, can greatly assist in the identification of subtle motifs.

Here we describe a new Gibbs strategy, called motif sampling, that addresses the problem of detecting motifs when little prior information about the number of occurrences of each motif is available. This is important because often some of the sequences under investigation may not contain the motifs common to the remaining sequences or the sequences may share varying numbers of repetitive motifs. In contrast to the site sampler, which iteratively samples sites for each motif, the motif sampler iteratively samples motif models (or possibly no model) for each site and thereby optimally partitions motif-encoding regions into different motifs. It can also be used to classify related motifs (and the proteins containing them). Another Gibbs sampling strategy (column sampling), which is applied within the motif sampling algorithm, optimizes motif lengths.

Porins are a major class of bacterial integral outer membrane proteins (iomps) that serve as diffusion channels for nutrients, waste products, and antibiotics (Nikaido, 1992, 1994; Cowan, 1993). X-ray and electron crystallographic analyses of four bacterial porins (the *Rhodobacter capsulatus* and *Rhodopseudomonas blastica* porins and *Escherichia coli* OmpF and PhoE) reveal that they form trimers of 16-stranded antiparallel β-barrels containing pores (Weiss et al., 1990; Jap et al., 1991; Cowan et al., 1992; Kreusch et al., 1994). There is evidence that other iomps, for example, OmpA and some specific uptake channels, also exist as β-barrels (Morona et al., 1984; Vogel & Jahnig, 1986; Nikaido, 1992 and references therein). In such proteins many of the β-strands that traverse the outer membrane would be expected to share similar environments (a hydrophilic pore on one side of the β-sheet and membrane phospholipids on the other); therefore, it is possible that many iomps share repetitive motifs corresponding to these strands. Previous predictions of β-strands in outer membrane proteins have been limited to specific families and have relied on global multiple sequence alignments and biochemical heuristics (Jeanteur et al., 1991, 1993; Schirmer & Cowan, 1993). Here motif sampling is used to automatically detect patterns conserved among very distantly related outer membrane proteins (a statistical significance test is used to distinguish these patterns from chance similarities). When applied to bacterial porins of known structure and related proteins, the sampler detected similar repeats that correspond to alternating membrane-spanning β-strands.

Results and discussion

Gibbs motif sampling for detecting multiple motifs and for detecting and classifying a single motif is illustrated using distantly related immunoglobulin fold proteins and helix-turn-helix DNA-binding proteins, respectively. It is then used to discover subtle, repetitive motifs in bacterial iomps.

Detecting multiple motifs: The immunoglobulin fold

The sampler's general applicability is illustrated by searching for multiple motifs in immunoglobulin fold proteins. The immunoglobulin fold is a structural domain present in many sequences including proteins that function in the immune system and cell-cell recognition, in several types of receptor proteins, and in other proteins with various functions (Hunkapiller & Hood, 1986; Williams & Barclay, 1988; Kuma et al., 1991; Jones, 1993; Bork et al., 1994; Harpaz & Chothia, 1994). It consists of about 100 residues forming two sets of antiparallel β-strands usually stabilized by a disulfide bond. Members of the immunoglobulin superfamily have been assigned to four different sets, V, C1, C2 (Williams & Barclay, 1988), and I (Harpaz & Chothia, 1994), having several distinguishing features yet also sharing some structural and sequence similarities. This superfamily provides an excellent test of the motif sampling algorithm because the proteins are highly diverse and contain variable numbers of immunoglobulin domains (from one to four or more).

Proteins from the immunoglobulin fold superfamily (258 sequences) were retrieved from the SwissProt database (version 29) (Bairoch & Boeckmann, 1992) and, in order to devise a stringent test set, similar sequences were removed using PURGE with an MSP cutoff score of 60 (see Methods), thereby leaving a set of 47 distantly related proteins (with an average pairwise MSP score of 35). Three motifs were specified for the search. The sampler converged on alignments of 66, 35, and 63 segments in 32, 18, and 34 sequences, respectively (Fig. 1). These correspond to the A′-B, C, and E-F β-strands of the immunoglobulin fold that were previously detected by a combination of pairwise alignment and visual inspection (Williams & Barclay, 1988; Harpaz & Chothia, 1994) and to conserved segments observed in V- and C2-type domains by Kuma et al. (1991).

The alignments from these motifs were used to search the SwissProt database for additional (unknown) members of the immunoglobulin superfamily using SCAN with the order option (see Methods). Two viral proteins (VGL2_EBV and YF30_FOWP1) had highly significant matches to the motifs ($P = 0.00001$ and 0.0000002, respectively) (Fig. 2A). Neither VGL2_EBV, which is a probable membrane glycoprotein (Mackett et al., 1990), nor YF30_FOWP1, whose function is unknown, had significant BLAST matches ($P \leq 0.01$ using a blosum62 scoring matrix) to any protein with Ig-like domains in the NCBI nonredundant database. Another protein, the sodium channel β_1 subunit from rat (CINB_RAT), showed marginally significant similarity to the motifs ($P = 0.03$); the presence of an Ig-like domain was confirmed by further analysis (Fig. 2B), which revealed weak yet significant, nearly global similarity to one protein, myelin P0, which is postulated to be the closest relative to the ancestral gene for the immunoglobulin superfamily (Lemke et al., 1988; Williams & Barclay, 1988). Because all Ig-like domains appear to be involved in binding functions, it is worth noting that the sodium channel β_1 subunit seems to exert its effects through binding to the sodium channel α subunit (Bennet et al., 1993).

Motif classification: The hth motif

Proteins are classified at different levels of divergence (for example, into superfamilies, families, or subfamilies) depending on the amount of conserved sequence similarity. Similarly, a motif model that seeks to capture the distinguishing characteristics of a set of related sequences can be constructed at different levels of divergence with less stringent models corresponding to motif "superfamilies" and more stringent models corresponding to motif "families" or "subfamilies." The motif sampler can be used

motif A	site	prob.	protein
ESESLLKPLANVTLTCQAR	13	0.9345	A1BG_HUMAN
ESSQVLHPGNKVTLTCVAP	196	0.9990	
EFSPEPESGRALRLRCLAP	289	0.9395	
PPFGGSAPSERLELHVDGP	360	0.7645	
TWSGAVLAGRDAVLRCEGP	387	0.9990	
PHTFESELSDPVELLVAES	456	0.7165	
KKSEHGNEGDVGVLTCKSP	119	0.8020	BASI_CHICK
LVHMTVVSGSNVTLNISES	30	0.9165	BCM1_HUMAN
LSSVPSSAHGHLQLVCHVS	211	0.9665	CD12_MOUSE
QATLDVEAGEEAGLACRVK	266	0.9305	
PLVVKVEEGDNAVLQCLKG	23	0.9995	CD19_HUMAN
SQDLTMAPGSTLWLSCGVP	185	0.9840	
REVVLGKAGDAVELPCQTS	26	0.9145	CD4_CANFA
SNTFYAREGDQVEFSFPLS	216	0.7115	
PHCTTVPVGASVNITCSTS	33	0.9995	CD7_HUMAN
PKKMDAELGQKVDLVCEVL	38	0.9720	CD8A_MOUSE
PQGGTVKVGEDITFIAKVK	64	0.9540	CPSF_HUMAN
LEDTTDYCGERVELECEVS	347	0.9995	
LTDQTVNLGKEICLKCEIS	437	0.9640	
DNTVTVIAGNKLRLEIPIS	530	0.5725	
LVNRLCHSGYMATLNCSVR	1050	0.7245	
PSVVLASSHGVASFPCEYS	43	0.9130	CTL4_MOUSE
PNTALLNEGDRTELLCRYG	26	0.9950	FAS3_DROME
NREGYFNEGTEFRARCSVR	135	0.6155	
PQWINVLQEDSVTLTCRGT	56	0.9900	FCGC_HUMAN
TPHLEFQEGETIVLRCHSW	137	0.9790	
NIGYTLYSSKPVTITVQAP	199	0.8775	
LPQLFLKVGEPLWIRCKAV	257	0.9945	FLT3_HUMAN
EENVILEKPSHVELKCVYT	74	0.5520	GP70_MOUSE
KKSLIAYVGDSTVLKCVCQ	167	0.8000	
TPEVKVACSEDVDLPCTAP	20	0.9980	HB15_HUMAN
THEKTPIEGRPFQLDCVLP	125	0.9770	HEMO_HYACE
SKDMMAKAGDVTMIYCMYG	237	0.7460	
EKVIVVKQGQDVTIPCKVT	334	1.0000	
LDWYPDAPGEMVVLTCDTP	35	0.9995	I12B_HUMAN
KTSATVICRKNASISVRAQ	293	0.6375	
LSEPEVSEWTTVTVECEAP	129	0.9940	ICA1_CANFA
PKKLAVEPKGSLEVNCSTT	33	0.8240	ICA2_HUMAN
LQPTLVAVGKSFTIECRVP	119	1.0000	
NGTVTSLPGATVTLICPGK	32	0.9995	IL6R_RAT
SVGKTLSPGTQVTTCCNSS	233	0.9175	
PSTISAFEGTCVSIPCRFD	27	0.9940	MAGL_MOUSE
VVPPEVVAGTEVEVSCMVP	144	0.9950	
NSSVEAIEGSHVSLLCGAD	246	1.0000	
NGTVVAVEGETVSILCSTQ	332	0.9995	
LESHCAAARDTVQCLCVVK	417	0.9420	
PAREQLNLRESATITCLVT	274	0.9980	MUCB_HUMAN
DREIYGAVGSQVTLHCSFW	35	0.9870	MYP0_MOUSE
TQDERKLLHTTASLRCSLK	36	0.7960	OX2G_RAT
PAWLTVSEGANATFTCSLS	39	0.9995	PD1_MOUSE
LPDWTVQNGKNLTLQCFAD	42	0.9980	PEC1_HUMAN
LDKKEAIQGGIVRVNCSVP	137	0.9880	
SSFTHLDQGERLNLSCSIP	332	0.9895	
DAQFEVIKGQTIEVRCESI	416	0.6135	
PAVFKDNPTEDVEYQCVAD	461	0.8445	
LSSKVVESGEDIVLQCAVN	508	0.9960	
PEEVNSVEGNSVSITCYYP	7	0.9995	PIGR_HUMAN
TKVYTVDLGRTVTINCPFK	119	1.0000	
PELVYEDLRGSVTFHCALG	224	0.8000	
PGNVTAVLGETLKVPCHFP	449	0.9975	
VKQEWAEIGKNVSLECASE	30	0.9535	PTP6_DROME
KNNKNSGCRSPLTVHCSLG	1376	0.7275	
NHTMEVEIGKPASIACSAC	225	0.9885	ST2_MOUSE
PDGIVTSIGSNLTIACRVS	227	0.9955	VB16_VACCV
DPKINVTIGEPANITCTAV	259	0.9945	VB19_VACCC
PPLASSSLGATIRLSCTLS	26	0.9960	VPR1_MOUSE
• • ******* • **			

motif B	site	prob.	protein
ETPDFQLFKNGVAQ	33	0.9645	A1BG_HUMAN
SEDRIYWQKHDKVV	64	0.8650	B7_MOUSE
PKPRFSWLENGREL	172	0.9755	
ESVNYTWYGDKRPF	159	0.6595	BCM1_HUMAN
SFANISWSRTDSLI	35	0.8900	CD12_MOUSE
KPVWVMWMRGDQEQ	234	0.8740	
PGSEILWQHNDKNI	53	0.9775	CD3E_HUMAN
LLLVYYWSKNRKAQ	145	0.6020	
EAKNITWFKDGKMI	49	1.0000	CD3G_HUMAN
PKLEVKWNKNGQEL	278	0.9995	CPSF_HUMAN
DDAQVKWFKNGEEI	367	1.0000	
ENIPGKWTKNGLPV	456	0.7750	
PPPKAMWSRGDKAI	551	0.9495	
PRPELTWKKDGAEI	865	1.0000	
PKPKITWMKNKVAI	1071	1.0000	
PPPYFVGMGNGTQI	136	0.8795	CTL4_MOUSE
PPANISWYIDNMPA	157	0.9910	FAS3_DROME
PQPKIEWTIDGAIV	264	0.9970	
ESDSIQWFHNGNLI	78	0.9970	FCGC_HUMAN
PLVKVTFFQNGKSK	159	0.7185	
NLMNVTWKKDDEPL	98	0.9240	GP70_MOUSE
QGVKYSWKKDGKSY	53	0.6800	HEMO_HYACE
PKPLITWKKRLSGA	147	0.9680	
PAPNVVWSHNAKPL	355	0.9940	
PPPLLTWMRDGMVL	267	1.0000	MAGL_MOUSE
PDPILTIFKEKQIL	353	0.9975	
RQIEVSWLREGKQV	80	0.6550	MUCB_HUMAN
ADVFVQWMQRGQPL	297	0.9165	
EPLIVTWQKKKAVG	58	0.8295	OX2G_RAT
PAPAISWKGTGSGI	166	0.9925	
EFPEIIIQKDKAIV	266	0.9575	PEC1_HUMAN
PPANFTIQKEDTIV	353	0.9740	
WTAPVQWFKNCKAL	147	0.9910	ST2_MOUSE
FLADVLWQINKTVV	250	0.8155	
HYNNITWYKDNKEI	181	0.9915	VB19_VACCC
• * **********			

Fig. 1. Three motifs detected in the immunoglobulin family. Using three 12-column models, each with an expectation of 90 sites, the sampler converged on the three alignments shown. These sequences contain many low complexity regions that were masked prior to analysis using the method of Wootton and Federhen (1993). The predictive probabilities with which the sites match the motif are indicated. Asterisks (*) below the alignments denote columns selected by the column sampler (see Methods). Proteins are designated by their SwissProt identifiers. (*Continues on facing page.*)

to classify motifs in this way by choosing the appropriate parameter specifications.

As described in the Methods, the motif sampler uses the following search parameters: k, the number of motif models; $C_{i=1...k}$, the number of columns (i.e., the minimum motif width) for each of the k models; and $e_{i=1...k}$, the expected number of sites for each motif. The stringency of the search depends on the values chosen for these parameters. The construction of more general models (tending toward superfamilies) is favored by specifying a small number of models each with relatively few columns and a high number of expected sites (say $k = 1$, $C_1 = 12$, and $e_1 = 2$ or more per sequence). Conversely,

motif C	site	prob.	protein
FHLNAVALGDGGHYTCRY	243	0.9985	A1BG_HUMAN
FELHNISVADSANYSCVY	338	0.9930	
LELIFVGPQHAGNYRCRY	434	1.0000	
LTIQNIQYEDNGIYFCKQ	105	1.0000	B29_MOUSE
LIILGLVLSDRGTYSCVV	104	1.0000	B7_MOUSE
YTIEGKVEDHSGVYECIY	80	0.9990	BASI_CHICK
RILKLNIEQDMGDYSCNG	175	0.5200	
LYISKVQKEDNSTYIMRV	94	0.9855	BCM1_HUMAN
LFIFNVSQQMGGFYLCQP	82	1.0000	CD19_HUMAN
LLLPRATAQDAGKYYCHR	246	1.0000	
SLKEFSELEQSGYYVCYP	83	0.6910	CD3E_HUMAN
LVIKDLEVADSGIYFCDT	94	1.0000	CD4_CANFA
LSLSWPELQDGGTWTCII	177	0.9800	
ITMHRLQLSDTGTYTCQA	99	1.0000	CD7_HUMAN
LTLNKFSKENEGYYFCSV	114	1.0000	CD8A_MOUSE
LSIMNVKPEDSDFYFCAT	102	0.9995	CD8B_MOUSE
MQIIKAKDNFAGNYRCEV	126	0.9980	CPSF_HUMAN
LNIDNCQMTDDSEYYVTA	308	0.5385	
LIIEGATKADAADYSVMT	398	0.9980	
LVIDHALTEDEGDYVFAP	486	0.9995	
LVIDIAERDDSGVYHINL	582	1.0000	
IFIRKAERSHSGKYDLQV	894	0.9995	
LEIGKPSPYDGGTYCCKA	1101	1.0000	
LTIQGLRAVDTGLYLCKV	114	1.0000	CTL4_MOUSE
VSIERVKASNNGQVKCSL	87	0.9175	FAS3_DROME
SYRFKANNNDSGEYTCQT	98	0.9905	FCGC_HUMAN
FSIPQANHSHSGDYHCTG	181	0.9995	
MVILKMTETQAGEYLLFI	128	0.6250	FLT3_HUMAN
AFVSSVARNDTGYYTCSS	315	0.8325	
LKIKHLLEEDGGSYWCRA	225	1.0000	GP70_MOUSE
LKIRNTTSCNSGTYRCTL	92	1.0000	HB15_HUMAN
ITIKSLTARDAGTYVCAF	88	1.0000	HEMA_VACCC
LVFLRPQASDEGHYQCFA	82	1.0000	HEMO_HYACE
YEIKGVTKDNSGYKGEPV	215	0.5500	
LLFKTTLPEDEGVYTCEV	290	1.0000	
LVIKGVKNGDKGYYGCRA	380	1.0000	
LTIQVKEFGDAGQYTCHK	75	1.0000	I12B_HUMAN
LVLRAVQVNDTGHYLCFL	77	1.0000	IL6R_RAT
LLLSTLSPELGGKYYFRG	102	0.8935	MAGL_MOUSE
LDLEEVTPGEDGVYACLA	290	0.9995	
LELPAVTPEDDGEYWCVA	377	1.0000	
IVIHNLDYSDNGTFTCDV	112	0.9980	MYP0_MOUSE
ITFWNTTLDDEGCYMCLF	106	1.0000	OX2G_RAT
MNILDTRRNDSGIYLCGA	108	1.0000	PD1_MOUSE
YFIPEVRIYDSGTYKCTV	94	1.0000	PEC1_HUMAN
VYSVMAMVEHSGNYTCXX	289	0.6790	
DFTKIASKSDSGTYICTA	371	1.0000	
WTKQKASKEQEGEYYCTA	557	0.9915	
VNIAQLSQDDSGRYKCGL	77	1.0000	PIGR_HUMAN
VVINQLRLSDAGQYLCQA	187	1.0000	
VVITGLRKEDAGRYLCGA	292	1.0000	
VILNQLTSRDAGFYWCLT	408	0.9995	
LTLNLVTRADEGWYWCGV	511	1.0000	
LTLLDVNINDSGNYTCTA	97	1.0000	PTP6_DROME
LEFTEVYKKENGTYKCTV	199	0.9905	
LKFLPARVEDSGIYACVI	78	1.0000	ST2_MOUSE
LFIDNVTHDDEGDYTCQF	172	1.0000	
LTLANFTTKDEGDYFCEL	90	1.0000	THY1_MOUSE
MLILNPTQSDSGIYICIT	84	1.0000	VB16_VACCV
ITIEDVRKNDAGYYTCVL	179	1.0000	
LNINPVKEEDATTFTCMA	294	0.9985	
LIIHNPELEDSGRYDCYV	208	1.0000	VB19_VACCC
LSISELQPEDEAVYYCAV	100	1.0000	VPR1_MOUSE
*** ** **** * * *			

Fig. 1. *Continued.*

more specific models (tending toward subfamilies) are favored when the number of motif models and the number of columns are increased and the expected number of sites is decreased. It is important to stress, however, that (due to its Bayesian statistical basis) the sampler will only subclassify a motif when warranted by the data (despite the parameter settings selected).

Motif classification is illustrated using the hth motif (Brennan & Matthews, 1989; Pabo & Sauer, 1992; Treisman et al., 1992), which is present in many DNA-binding proteins including the XylS/AraC (Gallegos et al., 1993), GalR/LacI (Weickert & Adhya, 1992), LuxR (Stout et al., 1991), and LysR (Viale et al., 1991) families. A diverse set of 90 known and putative hth proteins was selected from the SwissProt (version 30), PIR (release 42) (Barker et al., 1993), and GenBank (release 85) (Benson et al., 1993) databases. When the motif sampler was run on this set using low stringency parameter settings ($k = 1$, $C_1 = 12$, and $e_1 = 150$), 100 sites in 84 of the 90 sequences were detected (Fig. 3). On the other hand, when the parameters were set to more stringent settings ($k = 3$, $C_{i=1\ldots3} = 18$, and $e_{i=1\ldots3} = 20$) three distinct hth submotifs were detected: 17 sites in 17 proteins from the luxR family, 18 sites in 18 proteins from the lysR family, and 47 sites in 44 sequences from several other hth protein families (Fig. 3). Notably, the *Bacillus subtilis* CitR protein (BSCITRA_1) (Jin & Sonenshein, 1994) appears to contain two distinct types of motifs: a LysR-family motif near the N-terminus and a "multiple family" motif near the C-terminus.

A repetitive motif in bacterial iomps

Thirty-two bacterial iomps, which are or might be involved in substrate uptake, were selected from the SwissProt, PIR, and GenBank databases for analysis. These particular proteins were chosen because they constitute an extremely diverse set sharing no significant pairwise similarity; BLAST (Altschul et al., 1990) using a blosum62 scoring matrix (Henikoff & Henikoff, 1992) is unable to detect significant similarity ($P \leq 0.01$) between any of these sequences. Using an 11-column model (about 11 residues are needed to span the outer membrane), the sampler consistently converges on an alignment of about 130 segments (Fig. 4), with the number of repeats detected in individual proteins varying from one to nine. Note that the column sampler did not select a longer motif width, but maintained a contiguous motif model of 11 residues consistent with the length of the membrane-spanning β-strands.

By the Wilcoxon statistical test (see Methods), the repeats present in these sequences (designated as the iomp motif) are clearly significant ($P < 0.000001$); nevertheless, due to their subtle nature, it is difficult to decide whether or not certain sites actually match the motif. For this reason, it is helpful to consider the predictive probability of matching the motif returned by the sampler for each site (see Fig. 4 and Methods). Although by default only those sites with matching probabilities ≥0.5 are included in the alignment, it is sometimes informative to also examine sites with probabilities somewhat less than 0.5 (as is illustrated in Fig. 6) by using a program option.

The model obtained by the sampler suggests possible structural features of the corresponding protein regions. The alternating pattern of hydrophobic and hydrophilic residues (Table 1A) is characteristic of amphipathic β-strands. Consistent with this, the three repeats detected in the POR_RHOCA protein (Fig. 4), whose structure is known, correspond to membrane-spanning β-strands. This relationship is explored further through analysis of several porins of known structure and related sequences (see below). The predominance of aromatic residues near one end of the motif is also characteristic of membrane-spanning β-strands; aromatic residues have been observed to flank membrane-spanning segments in a number of proteins (including several

A

	motif A		motif B		motif C		protein
33	PPYVLVPEGSDINLTCFIK	51..59	DKVIVAWQQGDGEV	72..91	LHISKVSKEAEGSYMCVV	108	YF30_FOWP1
25	FENVTAHAGARVNLTCSVP	43..84	YSLTLEWVNDSNTS	97..100	LIIPNVTLAHAGYYTCNV	117	VGL2_EBV

B

```
                          10        20        30        40        50
CINB              MGTLLALVVGAVLVSSAWGGCVEVDSETEAVYGMTFKILCISCKRRSETT
                     : ::. .....:.. .  : .:.:. .. :   .. :   ..   ..
MYPO    MAPGAPSSSPSPILAALLFSSLVLSPTLAIVVYTDREVYGAVGSQVTLHCSFWSSEWVSD
                10        20        30        40        50        60

                60        70        80        90       100       110
CINB    AETFTEWTFRQKGTEEFVKILRYENEVLQLEEDERFEGRVVWNGSRGTKDLQDLSIFITN
        . .:: : ....:. . ..:.:....   ..: . :.:.  : :... :   : :: : :
MYPO    DISFT-WRYQPEGGRDAISIFHYAKGQPYIDEVGTFKERIQWVGDPSWK---DGSIVIHN
                 70        80        90       100          110

               120       130       140       150       160       170
CINB    VTYNHSGDYECHVYRLLFFDNYEHNTSVVKKIHLEVVDKANRDMASIVSEIMMYVLIVVL
        ..:...:...:.: .    :  .... :.  .  .: .. .  .......:.  ::...:
MYPO    LDYSDNGTFTCDVKNPP--DIVGKTSQVTLYVFEKVPTRYGVVLGAVIGGILGVVLLLLL
         120       130         140       150       160       170

               180       190       200       210
CINB    TIWLVAEMVYCYKKIAAATEAAAQENASEYLAITSESKENCTGVQVAE
         ..:.    ::. . .::
MYPO    LFYLI---RYCWLRRQAALQRRLSAMEKGKFHKSSKDSSKRGRQTPVLYAMLDHSRSTKA
              180       190       200       210       220       230
```

Fig. 2. Detection of putative Ig-like domains. **A:** Ig-like regions detected by SCAN in two viral proteins. **B:** Alignment of the rat sodium channel β_1 subunit (CINB_RAT) with the rat myelin P0 protein (MYP0_RAT). A potential CINB_RAT Ig-like domain (detected by SCAN as the motifs in bold; see Results) was confirmed by the following analysis. A BLASTP search of the NCBI nonredundant database with CINB_RAT as the query yielded a significant match ($P = 0.0024$) to but one protein, myelin P0 from horn shark (MYP0_HETFR); the rat homolog was used for the Smith and Waterman (1981) alignment shown. The alignment score of 161 was 16 standard deviations above the mean score for 10,000 alignments of shuffled sequences using the rdf2 program (Pearson & Lipman, 1988).

of the bacterial porins) where they are postulated to position the protein with respect to the lipid bilayer (Cowan, 1993).

The iomp repeats are similar to a conserved C-terminal outer membrane protein pattern described by Struyve et al. (1991). In fact, 14 such C-terminal patterns were included in the alignment detected by the sampler (Fig. 4). Thus, it appears that a pattern like that of Struyve et al. (1991) is also present at many internal locations in outer membrane proteins. These bacterial iomp repeats also show significant similarity to regions in several mitochondrial porins (to be described elsewhere; Mannella et al., 1996).

Iomp repeats in other bacterial membrane proteins

A SCAN search was performed on a set of 65 bacterial iomps from the SwissProt database having functions apparently unrelated to substrate uptake (and that were consequently not included in our initial set) in order to detect any additional proteins with the iomp motif. Two secreted proteases (OMPT_ECOLI, and OMPP_ECOLI) and a protein involved in the export and assembly of fimbrial subunits across the outer membrane (FAND_ECOLI) yielded matches that are significant at the $P \leq 0.02$ level (Table 2). Assuming that these matches are biologically significant, what function might they imply? IgA-specific serine protease from *Neisseria gonorrhoeae* has a C-terminal helper domain that associates with the outer membrane to form a pore for excretion of the protease domain (Pohlner et al., 1987). By analogy, perhaps the repeat regions correspond to pore-forming β-strands involved in excretion of these proteins across the outer membrane. Notably, the OmpP repeats are located in the C-terminal half of the protein, which is protected from proteinase K digestion in intact cells (Kaufmann et al., 1994) (implying that it is in the membrane).

Repeats in bacterial porins of known structure

The motif model obtained from the extremely diverse initial set of bacterial outer membrane proteins is likely to be highly generalized, and, conversely, repeats present in specific subfamilies of the outer membrane proteins are likely to share additional discriminating features. Therefore, in order to better characterize the repeats present in bacterial proteins that are related to the porins of known structure, a subset of the iomps was obtained and analyzed as follows.

Table 1. *Outer membrane protein motif models*[a]

Column	Residue frequency C	G	S	T	N	D	E	Q	K	R	H	W	Y	F	V	I	L	M	A	P	Information (bits)
A. Motif model for the outer membrane protein alignment in Figure 4																					
1	.	.	.	.	13	22	.	.	15	9	.	6	.	.	.	.	.	.	.	.	0.7
2	.	.	.	10	14	.	.	8	.	.	.	.	.	.	.	.	.	.	14	.	0.3
3	.	.	.	.	.	.	.	.	.	.	5	16	16	11	15	.	.	.	.	.	0.9
4	.	.	.	11	.	.	.	.	.	.	5	.	16	.	.	.	.	.	.	.	0.4
5	.	.	.	.	.	.	.	.	.	.	.	.	.	13	21	7	22	.	14	.	0.8
6	.	30	19	.	.	.	.	.	.	15	.	.	.	.	.	.	.	.	.	.	0.7
7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	21	.	30	.	19	.	0.9
8	.	38	.	.	15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0.8
9	.	.	.	.	.	.	.	.	.	.	.	.	28	7	23	.	11	5	.	.	1.0
10	.	.	.	.	.	13	.	.	10	17	.	.	.	.	.	.	.	.	.	.	0.4
11	.	.	.	.	.	.	.	.	.	.	.	.	42	27	.	.	.	.	.	.	1.6
B. Motif model for the porin-like protein alignment in Figure 5																					
1	.	.	18	.	10	15	.	.	13	7	.	.	.	.	.	.	.	.	.	.	0.5
(2)																					–
3	.	.	.	20	12	21	8	7	.	.	.	.	.	.	.	.	.	.	.	.	0.7
4	.	.	11	.	.	.	12	.	.	.	.	.	.	.	.	.	.	.	18	.	0.4
5	.	.	.	.	.	.	.	.	.	.	.	11	11	13	21	.	16	.	.	.	1.0
(6)																					–
7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	34	.	32	.	.	.	1.2
8	.	64	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.2
9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	13	7	22	6	31	.	0.9
10	.	.	11	10	10	.	.	10	.	16	.	.	.	.	.	.	.	.	.	.	0.5
11	.	.	.	.	.	.	.	.	.	.	9	.	58	.	.	.	.	.	.	.	1.9
12	.	.	.	.	.	22	.	12	13	8	.	.	.	.	.	.	.	.	.	.	0.6
13	.	.	.	.	.	.	.	.	.	.	.	.	.	52	.	.	11	.	.	.	1.6

[a] Model target frequencies are shown (as percentages) for residues with elevated frequencies. Columns 2 and 6 in Table 1B were deselected by the column sampler (see Methods).

A set of bacterial iomps, consisting of the 32 proteins analyzed above and related proteins, was searched for sequences having at least marginally significant BLASTP matches ($P \leq 0.05$) with one or more of the porins of known structure. This yielded a set of 25 proteins from which closely related sequences were removed using PURGE with a cutoff of 200 (because the *E. coli* porins with known structure [OmpF and PhoE] are closely related, only OmpF was retained). When the motif sampler was applied to the remaining set of 19 sequences, an optimum alignment consisting of 70 segments with a motif width of 13 was found (Fig. 5).

Table 2. *Repeats detected in the Escherichia coli OmpT, OmpP, and FanD proteins*[a]

Segment	Site	Protein (*P*-value)
PNYRLGLMAGY	144	OMPT_ECOLI (0.013)
EDFELGGTFKY	210	
NYYSVAVNAGY	249	
YNFITTAGLKY	305	
DNSVTGANVSY	488	FAND_ECOLI (0.015)
RQFYSNSGVTY	538	
DNESVSLSTNY	579	
DNFEFGGAFKY	210	OMPP_ECOLI (0.020)
NYYSVAVNAGY	247	
YNFITTAGLKY	303	

[a] SCAN program parameters used were $R_{\min} = 3$, and $R_{\max} = 6$.

For each of the three porins with known structure, four repeats were detected having several notable characteristics. They correspond to alternating membrane-spanning β-strands, which are oriented with their N-terminal ends outside of the bacterial cell and their C-terminal ends on the periplasmic side of the membrane (Fig. 6). These strands occur on the membrane interface (as opposed to the trimeric interface) of the porin β-barrel (Fig. 6B). Comparison of the iomp motif model with this porin motif model (Table 1) reveals several similarities (e.g., alternating amphipathicity, and a predominance of aromatic residues near one end); the differences presumably reflect underlying functional distinctions.

Although these porin repeats may only function in pore formation and retention of the protein within the outer membrane, it is tempting to consider whether they might be involved in membrane insertion. This is suggested by the fact that deletion of the C-terminal segment of *E. coli* PhoE (which contains this motif) completely prevents incorporation of the protein into the outer membrane (Bosch et al., 1989; Struyve et al., 1991). Insertion of a β-barrel structure into a membrane may be more difficult than insertion of a protein containing one or more hydrophobic membrane-spanning α-helices. In an α-helix, the

A

segment	site	prob.	protein
LVVFNQLLVDRRVS**ITAENLGLTQPAVSNAL**KRLRTSLQDPLFVRTHQGMEPT	11	1.0000	NAHR_PSEPU
LEYFYQLSKLRSFT**NVAKHFRVSQPTISYAI**KRLETYYDCDLFYKDSSHQVVD	8	1.0000	MLER_LACLA
LQALDAVIRERGFE**RAAQKLCITQSAVSQRI**KQLENMFGQPLLVRTVPPRPTE	9	1.0000	ICIA_ECOLI
LRALCAIADAGSLH**RAARRLGVAQPTLSTQL**TRIEQALGGPLFTRERTGCRPT	8	1.0000	A48990
LYYFWHVYKEGSVV**GAAEALYLTPQTITGQI**RALEDALQAKLFKRKGTWSRTQ	11	1.0000	NHAR_ECOLI
LQAALRVAETGSFQ**EAAQKVGCNQSTISRQV**KGLEDELGIALFRRQGRMKLTA	6	1.0000	SYONIRB_3
LRMLVMIEEHGQVS**AAAAAMNMTQPAASRML**SEMEAIVKSPLCQRASRGVVLT	4	1.0000	GBPR_AGRTU
LKIFITLMETGSFS**IATSVLYITRTPLSRVI**SDLERELKQRLFIRKNGTLIPT	9	1.0000	VRPR_SALTY
VKAFHALCQHKSLT**AAAKALEQPKSTLSRRL**AQLEEDLGQSLLMRQGNRLTLT	7	1.0000	IRGB_VIBCH
LEVVDAVARNGSFS**AAAQELHRVPSAVSYTV**RQLEEWLAVPLFERRHRDVELT	7	1.0000	YDHB_ECOLI
LHTFVTAAKYENFR**KTAETLFLSQPTVTVHI**KQLEKEISCNVFDVKGRQIQLT	6	1.0000	BSCITRA_1
CRAFVKVSERGSFTVGAAAAQMSQSVASRRVAALEKHFGERLFDRASRRPSLT	7	1.0000	BLAA_STRCI
WMIFIKVAEVGNLS**RAARELDISISAVSKSL**SRLENSIEVTLLRRDSHHLELT	13	1.0000	SINR_SALTY
LRVVAAINRCGSFN**RAAKMLNVEETTIARRL**ARLEGSLGCVLFQAVDGQRRPT	6	1.0000	PDU12464_2
LKPYWCSAKEKTMS**RTGSQLYISQSAVSKRI**ANLEKKLSKKLIVPAGRHIKLT	185	1.0000	YREC_VIBCH
RYIVEVVNHNLNVS**STAEGLYTSQPGISKQV**RMLEDELGIQIFSRSGKHLTQV	7	1.0000	CYSB_ECOLI
LLRSFVVIAEVRAL**SAAARVGRTQSALSQQM**KRLEDIVDQPLLPAHRPRRGAD	18	1.0000	DGDR_BURCE
LKIISVIAASENIS**HAATVLGIAQANVSKYL**ADFESKVGLKVFDRTTRQLMLT	11	1.0000	YIAU_ECOLI
* * ** ● ***** * ** * ** * *			

B

segment	site	prob.	protein
LTKREKECLAWASEGKSTW**DISKILGCSERTVTFHL**TNTQMKLNTTN	183	1.0000	LUXS_VIBFI
LTKREREVFELLVQDKTTK**EIASELFISEKTVRNHI**SNAMQKLGVKG	12	1.0000	GERE_BACSU
LTPRECLILQEVEKGFTNQ**EIADALHLSKRSIEYSL**TSIFNKLNVGS	154	1.0000	COM1_BACSU
LTAKEREIVGMVREGASNK**LIARQLDISLSTVKTHL**RNIFAKTEVVN	162	1.0000	KPNMOAR_1
LSPREQAVMKLVATGLMNK**QVAAELGLAEITVKIYR**GHVMKKMRARS	158	1.0000	NODW_BRAJA
LTRRERQVAELLLQGLDTE**AIAAALGIGNGTVKNHR**KHLYGKLRLGS	124	1.0000	ASEXAN2_1
LSQTESNMLKMWMSGHDTI**QISDKMQIKAKTVSSHK**GNIKRKIKTHN	142	1.0000	RCSA_ERWST
LTQRQYEILVLLSRGHPVK**TISRMLGISEATTKAHI**NALYRRLEVRS	153	1.0000	PSEEPSR_1
LTGREEEILGMITEGMSYR**DIADRACISYKTVSNVS**LVLKDKLGAAN	162	1.0000	MOXX_PARDE
LSPRELSVLSMAEGGDTVA**GIAGRLHLTPGTVRNYL**AAAIRKSGARN	141	1.0000	A47096
LSPKESEVLRLFAEGFLVT**EIAKKLNRSIKTISSQK**KSAMMKLGVEN	151	1.0000	RCSB_ECOLI
VSDREVIILRLLANGMKDV**AMARSLGISTRTLRRVI**TDLMGKLGVSS	191	1.0000	BRPA_STRHY
LTDAELRVAALAADGMANR**AIAAELQVTLRTVELHL**TKAYRKLGIRG	817	1.0000	STMCHO_2
LTANERNVLAMLVKGMDIR**QISCELNVHLKTIYSVR**YHVLTKLGCRT	116	1.0000	S25253
LTQKEQAVLQCLLKNGGIN**EIKSQLKIEEKTLSCYR**SKITRKFGCKR	66	1.0000	S06971
LDPKEATYLRWIAVGKTME**EIADVEEVKYNSVRVKL**REAMKRFDVRS	174	1.0000	TRAR_AGRVI
LTDKEFETLVLYCQMMNVQMVADYQNRKPDVIIKHLKSCRQKIGVES	19	1.0000	TRJ8_ECOLI
** ** * * * ** * * ** * ** * *			

Fig. 3. Classification of hth DNA-binding proteins. Putative hth regions were classified by the motif sampler using high stringency parameters. **A:** LysR family proteins. **B:** LuxR family proteins. **C:** Regions from other hth families that were classified as a single group by the sampler. **D:** Regions detected with low but not with high stringency parameters. Regions detected under low stringency that were also detected under high stringency are red in A, B, and C. (*Continues on facing page.*)

polarity of the peptide backbone is neutralized by hydrogen bonds internal to the helix, whereas for β-strands, the backbone dipolar moments are not neutralized until the strands become hydrogen bonded to adjacent strands. Therefore, if porin insertion requires specific facilitating factors, then perhaps these repeats serve as recognition signals for processive insertion of pairs of β-strands (one for each conserved repeat) into the membrane.

Conclusion

The selection of a particular sequence analysis method depends on the nature of the similarities one is attempting to detect and on the availability of relevant sequence data. The motif sampler addresses the problem of detecting subtle similarities in a relatively large, diverse set of related sequences. How does it differ from other methods (including the site sampler) and under what circumstances is it to be preferred?

Lawrence et al. (1993) have compared Gibbs sampling with several other motif methods and this need not be reiterated here. More recently, however, some closely related methods that utilize HMM for multiple sequence alignment have been described (Baldi et al., 1994; Krogh et al., 1994). Like Gibbs sampling, the HMM methods utilize one-to-many sequence comparisons in conjunction with an iterative procedure that eventually converges on an optimum alignment. Unlike the Gibbs methods, which are stochastically based, the HMM methods are EM based and consequently are more likely to get trapped in local optima. And, as currently implemented, their main application has been for gap-based global alignment of relatively closely related sequences. Because the Gibbs methods have been applied to detecting subtle block-based motifs, a more extensive comparison is not appropriate at this time. Nevertheless, as both the Gibbs and HMM methods are developed further, there is great potential for cross fertilization of ideas between the two approaches.

C

segment	site	prob.	protein
SQK**ETGDILGISQMHVSRLQ**RKA	223	0.9945	RPSB_BACSU
GTE**KTAEAVGVDKSQISRWK**RDW	25	0.7465	RPC2_LAMBD
TRQ**EIGQIVGCSRETVGRIL**KML	169	1.0000	CRP_ECOLI
TQR**AVAKALGISDAAVSQWK**EVI	12	0.9995	RCRO_BPP22
HLK**DAAALLGVSEMTIRRDL**NNH	23	1.0000	DEOR_ECOLI
TQR**EIAKELGISRSYVSRIE**KRA	250	1.0000	RPSK_BACSU
TTR**KLAQKLGVEQPTLYWHV**KNK	26	0.9605	TER2_ECOLI
SQRELKNELGAGIATITRGSNSL	66	0.9600	TRPR_ECOLI
TRG**DIGNYLGLTVETISRLL**GRF	196	1.0000	FNR_ECOLI
EKE**EVAKKCGITPLQVRVWF**INK	99	0.5625	MTA1_YEAST
TLA**IIADVFNVSEITIRKRL**ESE	181	0.9575	S20081
AVG**ALAHKVGLSQSALSQHL**SKL	48	0.9815	NOLR_RHIME
AYA**ELAKQFGVSPGTIHVRV**EKM	24	1.0000	ASNC_ECOLI
ALT**ELAQQAGLPNSTTHRLL**TTM	58	0.9635	ECOICLRA_2
TRL**DVADYLGMTIETVSRTI**TKL	178	1.0000	S28677
THQ**VIAELSGSTRVTTTRLL**GEF	155	1.0000	CYSR_SYNP7
AWT**QLADYLGTTPETVSRTL**KRL	171	1.0000	FLP_LACCA
TTE**ALSEQLKVSKETIRRDL**NEL	20	1.0000	FUCR_ECOLI
QVQ**DLAGVFAASEATIRADL**RFL	23	0.9005	GATR_ECOLI
TVE**KVVERLGISPATARRDI**NKL	21	1.0000	ECOUW93_103
AEQ**QLAARFEVNRHTLRRAI**DQL	37	0.9990	PHNF_ECOLI
AER**ELSELIGVTRTTLREVL**QRL	33	1.0000	S01288
SER**ELGELLGIKRMTLRQAL**LNL	32	0.9995	YIHL_ECOLI
SEN**ELAASMGVSRTPVRESL**ILL	33	0.9995	YIN1_STRAM
PQR**AIAEALGVDLTTVTRAL**NEA	38	1.0000	YRDX_RHOSH
PTR**MMAEDLGVSRNTVITTY**DAL	37	0.9185	S43169
QQGAILGYAGIDPKTMREGINSL	446	0.8130	S43169
SEN**TIAAEFSVSRSPVREAL**KIL	43	0.9975	GNTR_BACSU
SLH**DVARLAGVSKSTVSRVI**NDE	3	1.0000	SCRR_VIBAL
TIK**DIAELAGVSKATASLVL**NGR	7	1.0000	SCRR_KLEPN
RLA**QVAKKVGVSEATVSRVL**NGK	4	1.0000	S21353
TLK**ELAEAAGVSKATLHRFC**GTR	25	1.0000	NFXB_PSEAE
SLK**AIATTLGISVTTVSRAL**GGF	1	1.0000	RAFR_ECOLI
TRD**DVARLAGTSTAVVSYVI**NNG	5	0.9430	STMGLNR_2
TTR**EIAKATGTSLQTVITTL**KIL	78	1.0000	REMA_STAAU
SIT**EAAAALGVSRKTLSAIL**NGH	26	1.0000	YSY1_SYNP7
TFK**QIALESGLSTGTISSFI**NDK	20	1.0000	VPB_BPMU
TIR**EVAEGTGLSTATIERWT**SAP	31	1.0000	CGPXZ_4
SMR**AIAAEIGCSVGLVHRYV**KEV	77	0.9960	CGPXZ_4
VLH**DIAEAVGMHESTISRVT**TQK	391	0.9980	RP54_AZOVI
TQQ**ELADWQGVSVDTIRRVL**KNA	25	1.0000	VR2B_BPT4
TLQ**QVADASGMTKGYLSQLL**NAK	17	1.0000	NADR_ECOLI
TIR**ELADELGVSKQRIQQII**AKL	4	1.0000	PHREP_2
TKR**FIKEGLGVSFLPLSTVK**REL	226	0.9945	BSCITRA_1
TIG**VFAKAAGVNVETIRFYQ**RKG	9	0.9930	MERR_PSEAE
TPG**EVAKRSGVAVSALHFYE**SKG	13	0.9790	SOXR_ECOLI
AII**KIAQRIGIPLATIGEAF**GVL	58	0.7495	SOXR_ECOLI
* ***** ***** **** ** *			

D

segment	site	prob.	protein
EIAEFLDVSEEEVLETM	137	0.9850	RPSB_BACSU
KTAKDLGVYQSAINKAI	18	0.9695	RCRO_LAMBD
KAARLLGMTPRQVAYRI	498	0.9915	NIFA_KLEPN
EIAHALCLTERQIKIWF	329	0.9785	HMAN_DROME
SVAQHVCLSPSRLSHLF	199	0.9005	ARAC_ECOLI
SLAKALKISHVSVSQWE	25	0.9905	DICA_ECOLI
RAALMMGINRGTLRKKL	76	0.9940	FIS_ECOLI
DLLEHFQFSQPTLSHHM	34	0.7255	ARSR_STAAU
DIANILGVTIANASHHL	61	0.9950	CADC_STAAU
HVADALGITEGENVIHL	116	0.5245	PHNF_ECOLI
HIATLSKCSTPALRIAY	289	0.7405	YRDX_RHOSH
RAAQGQGVNFSPLSRQF	424	0.9650	S43169
STAEGGECSSTAIRAII	433	0.9245	RP54_AZOVI
KNAEEAKRPKVTISGDI	45	0.8595	VR2B_BPT4
DLAIVMDSPTLTTSAGL	141	0.5185	SYONIRB_3
GSLKNLIISALTISGQK	104	0.6405	VRPR_SALTY
RMSETLGISANHTEQTQ	210	0.6295	NODW_BRAJA
RISSGLDVHPLTLSQTE	130	0.9540	RCSA_ERWST
DIALLNYLSSVTLSPAD	198	0.7700	RCSB_ECOLI
ETADAIDVSDREVIILR	184	0.8885	BRPA_STRHY
KSAQRLGFSLDEIAELL	58	0.8450	MERR_PSEAE
RLALDAGVSVHIVRDYL	8	0.9745	MERD_PSEAE
**** **** *** *			

Fig. 3. *Continued.*

Neuwald and Green (1994) describe an efficient method to search exhaustively for statistically significant patterns and to assemble the corresponding alignments. When no prior information concerning the input sequences is available, this method is often preferable over the Gibbs methods because it does not require specification of the number of types of motifs, their minimum lengths, or estimates of the number of occurrences of each motif. Because it does not use a probabilistic motif model, however, it may have difficulty detecting weakly conserved regions that lack sufficient exact matches to specific patterns. When prior information concerning the input sequences is available, or when searching for specific types of motifs, the Gibbs methods are preferable because they can use this information to constrain the search and increase sensitivity.

The choice between the site and motif samplers depends on the amount of prior information available. When the number and distribution of the motif sites is uncertain (as was the case for the iomps), the motif sampler is preferable because it only requires a prior rough estimate of the number of occurrences of each motif in the entire sequence set. When there is reason to suspect a specific number of occurrences for each motif in each sequence, however, this greater flexibility will result in a loss of sensitivity for two reasons: (1) because many ways of distributing the motif sites among the sequences must be considered; and (2) because groups of closely related (i.e., highly correlated) motif sites are more likely to bias the model against distantly related sites. Therefore, in this case, the site sampler is preferable because it can take advantage of this prior infor-

segment	site	prob.	protein
GNYTVGLGYEK	159	0.9005	PORI_RHOCA
KAYGLSVDSTF	230	0.8795	
DVTYYGLGASY	259	0.6200	
NRIDLYTGYTY	107	0.7655	NGOOPC_1
LNFRVGAGLGF	125	0.7870	
SEVKFDLNSRY	178	0.7570	
NGWGFGLGANI	206	0.8395	
REYGLRVGIKF*	262	0.9290	
DVYYAGLNYKN	211	0.8315	NMPORAP15_1
DQIIAGVDYDF	340	0.8865	
NAASVGLRHKF*	376	0.8685	
SQWALRVKYNF	205	0.9455	S42207
DWLTVRPNLQY	420	0.8875	
INNGIQVGAKY	199	0.5285	OMP2_HAEIN
KIAYGRTNYKY	217	0.6775	
NGVLATLGYRF	238	0.9655	
KRYFVSPGFQY	273	0.7610	
KSVGVGLRVYF*	351	0.9240	
NDYGTSVNLGY	460	0.8455	HIU13961_1
DGVSLGGNVFF	478	0.8180	
NSYYVGLGHTY	521	0.9955	
KYYKLSADVQG	599	0.6320	
SRIRASTGVGF	751	0.5380	
WNGSVRGRVGY	113	0.7775	RLROPB_1
FGYTVGAGVEA	155	0.6960	
NNITTRLEYRY	169	0.9825	
NSVKLGIGVKF*	201	0.9345	
NTWYTGAKLGW	26	0.7430	OMPA_ECOLI
VNPYVGFEMGY	66	0.6485	
DNGMLSLGVSY	179	0.8835	
DWWHQSVNVVG	33	0.6080	TSX_ECOLI
KEWYFANNYIY	122	0.5995	
STWYMGLGTDI	143	0.8175	
DHWHYSVVARY	247	0.7980	
WGGYLVVGYNF*	284	0.8190	
DRPTFSAGAVY	68	0.8715	FADL_ECOLI
NAWSFGLGFNA	165	0.5855	
DAYRIALGTTY	350	0.6010	
DNWTFRTGIAF	364	0.9855	
KDASVDVGVSY	406	0.8600	
KAWLFGTNFNY	436	0.9330	
DTTYARLGFKG	55	0.6250	NMPC_ECOLI
DGFGFSATYEY	193	0.9540	
FDFGLRPSVAY	287	0.6820	
KYVDVGATYYF	315	0.6570	
DIVAVGLVYQF*	355	0.6715	

segment	site	prob.	protein
RQYYLNSNYTI	234	0.8055	PORD_PSEAE
KHHETNLEAKY	389	0.8655	
DQNEFRLIVDY	428	0.8950	
NVVHLGLQYAY	226	0.9820	PORP_PSEAE
DGLVMRLQYVF*	430	0.5525	
KSFYFDTNVAY	81	0.9680	LAMB_ECOLI
PGGTLELGVDY	208	0.7545	
KGLSQGSGVAF	269	0.5505	
WTVGIRPMYKW	331	0.5785	
DEWTFGAQMEI	434	0.8525	
AAYPLRLRYKF	25	0.6315	TP50_TREPA
RRKLASLGYQF	369	0.7925	FECA_ECOLI
SAHEVGVGYRY	433	0.9980	
GNWTITPGMRF	491	0.9505	
RTWELGTRYDD	571	0.6565	
DNVSIYASYAY	635	0.9335	
PKHKGTLGVDY	665	0.6080	
GNWTFNLNSDF	678	0.9705	
KEFHIEPLLRY	347	0.5355	VIUA_VIBCH
WNYEFYTRHRF	496	0.9475	
SYWVANAQLAY	629	0.6005	
KKVLVDANLGW	434	0.5600	OAR_MYXXA
NRVTLNLGVRY	623	0.9980	
DNVTVYLNRTF	827	0.8680	
DGWLAQANYTW	839	0.8980	
NALSASVGVSY	908	0.9090	
RQVRFGIRYTF*	1051	0.9285	
RSWLFRPGFRF	310	0.9835	TBP1_NEIGO
WADYARLSYDR	422	0.5340	
IRHNLSVNLGY	493	0.5235	
DYYYQSANRAY	514	0.6735	
RWADVGAGLRY	594	0.5680	
SRYVVGSGYDQ	789	0.5440	
KHFTLRAGVYN	858	0.7925	
RNYTFSLEMKF*	905	0.9885	
KNPMSGTGLRW	815	0.6240	NFRA_ECOLI
WPHKVSLGVEY	955	0.9365	
LYPYVGVGVGR	141	0.6080	ALKL_PSEOL
WAPAFQVGLRY	171	0.7985	
NSWMLNSDVRY	185	0.9545	
DPFILSLGASY	218	0.7485	
DTNAFSVGYAR	24	0.7875	PAGC_SALTY
FAWGAGVQMNP	147	0.5105	
NGFNVGVGYRF*	178	0.9595	

segment	site	prob.	protein
KSYGALLNFGY	54	0.6055	OMPV_VIBCH
NADLSGLNYRF	70	0.7610	
GTYLTGSGVAY	91	0.8245	
KGYKTGVNYFH	147	0.6135	
KAYHAGGDFSY	196	0.6785	
NQWLVGATVAY	245	0.9880	
DVAGFRAGLFY	127	0.8220	OM3A_RHILV
GTFYAGLSVDE	168	0.5140	
DAWKVGLTVDY	304	0.9780	
ENFYAKASVQY	318	0.6490	
DGVYMDVDAGY	485	0.7450	RIRTSS56A_1
NAFVASAGIRY	509	0.8270	
KNVSASVLFDF	164	0.7740	FNOMPI_1
EKFGLRPQYKY	187	0.7730	
NQYHLGFESDF	208	0.8325	
LNFALNLEYDF	222	0.9015	
GGARVEVEVGY	126	0.5295	AMU07862_1
FAYRVKAGLSY	335	0.7120	
FGGELGVRFAF*	399	0.5060	
DTPLTRVTVDY	220	0.9080	YEFCUA_1
DQFGVRVNVLH	250	0.6345	
RTTAVSTGLDY	273	0.6405	
DRARTSLDVGY	286	0.8140	
WTVYGSVGASR	352	0.5290	
ITHKVNLGYAA	410	0.6300	
DKVSLMLGVRR	481	0.6145	
PWTRLDLGVRY	698	0.8955	
RALKLSVSMDF*	748	0.7975	
GTWGIRAGQQF	61	0.7145	PSEOPRH1_1
KNASIEGGYRY	154	0.9055	
SQFYLGANYKF*	190	0.9945	
GQWYLGVDANG	74	0.7965	FopA
RNVQASVDYRY	170	0.9700	
DDVTVYLDTKF	272	0.5460	
DAISIRAGYYG	56	0.6060	OMP1_CHLPS
WQVGLALSYRL	278	0.6210	
RAAHMNAQFRF*	392	0.6060	
SPYYVQADLAY	31	0.7340	OPR1_NEIME
IHPRVSVGYDF	76	0.9570	
SSLGLSAIYDF	136	0.6805	
FKPYIGARVAY	152	0.8815	
PKLTLDTGYRY	226	0.8990	
KTHEASLGMRY	248	0.9930	

Fig. 4. Motif detected among bacterial iomps. Thirty-two distantly related bacterial iomps (see text) were searched for a single motif (11-column model) with a prior expectation of 130 repeats. C-termini are indicated with asterisks. The alignment is the optimal of 300 independent runs. Proteins are designated by their SwissProt, PIR, or GenBank identifiers, except for FopA, which is taken from Leslie et al. (1993).

mation to decrease the uncertainty about the alignment and consequently yield greater sensitivity.

The original site sampler used an information per parameter (ipp) (Lawrence et al., 1993) criterion for determining optimum motif width. For statistical reasons, it could not be used for the motif sampler. Even for the site sampler, however, it does not necessarily yield optimum results. For example, the ipp values for the optimum pattern widths of 18–21 residues reported by Lawrence et al. (1993) for a set of 30 hth proteins can be exceeded by using a (biologically unrealistic) pattern width of 3. Width optimization by column sampling (which has also been added to the site sampler) avoids these problems and the need to perform multiple runs using different model widths.

Different methods often have complementary strengths, and the choice of which method to use depends on the nature of the search being conducted. A useful strategy for detecting motifs in uncharacterized protein families (where prior information is minimal) is to first perform a very broad search using, for example, the method of Neuwald and Green (1994) to get an initial idea about the numbers and types of motifs present. Then a more specific search can be performed that, depending on the nature of these motifs, uses one of the Gibbs samplers. Finally, the SCAN program can be used to search for other proteins matching the motifs found. Application of the appropriate tools in this way is useful for probing the relationships between distantly related sequences, which, in turn, helps to highlight key structural and mechanistic features important to protein function.

Methods

The statistical basis for the motif and column sampling algorithms and the Wilcoxon test is given in Liu et al. (1995).

Motif sampling

The motif sampler partitions the input sequence into regions corresponding to a specified number of motif models, including a "null" model representing those regions that contain no motifs. To accomplish this, it maintains two evolving data structures for each of the (non-null) motifs: (1) an alignment of sequence segments; and (2) a corresponding residue frequency model consisting of target probabilities obtained from the observed

segment	site	prob.	protein
DAQEMAVAAAYTF	146	0.9605	PORI_RHOCA
DMEQLELAAIAKF	180	0.8575	
DVTYYGLGASYDL	259	0.9960	
SDMVADLGVKFKF	289	0.5690	
KAEQWATGLKYDA	232	0.9650	OMPF_ECOLI
KTQDVLLVAQYQF	275	0.9955	
LVNYFEVGATYYF	313	0.9910	
SDDTVAVGIVYQF	350	0.9980	
DNDIAFVGAAYKF	190	0.9980	PORI_RHOBL
AGDQVTLYGNYAF	220	0.7770	
ADTAYGIGADYQF	249	0.9635	
NETVADVGVRFDF	277	0.9650	
SSAEWAVAAEYAI	266	0.8215	OM3A_RHILV
LGDAWKVGLTVDY	302	0.9850	
SYSGFQFGIGYSF	174	0.7225	S16480
TPRSYGLGGSYDF	244	0.8195	
KANSYMVGLSAPI	303	0.5860	
KMNVFSLGYTYDL	338	0.9880	
KSTAVGVGIRHRF	373	0.9995	
SNTTWSLAAAYTL	273	0.9840	PORD_PSEAE
DQNEFRLIVDYPL	428	0.5100	
LVEGLNFALQYQG	162	0.7020	S34263
NGDGFGMSTSYDF	199	0.7555	
TAEAWTIGAKYDA	250	0.9065	
KTQNFEVVAQYQF	294	0.9840	
LVKYVDVGMTYYF	341	0.9660	
TDDIVGVGLVYQF	382	0.9980	
GLDGLVLGANYLL	164	0.8310	HIMOMP2B_1
ISNGVQVGAKYDA	200	0.9305	
REQAVLFGVDHKL	334	0.8370	
KEKSVGVGLRVYF	374	0.9725	
NDTITVVGAQETF	65	0.6635	S30948
RTTAVSTGLDYRG	273	0.7340	
PWTRLDLGVRYTM	698	0.9000	
DPRALKLSVSMDF	746	0.6545	
* *** *******			

segment	site	prob.	protein
DFNADLSGLNYRF	68	0.7890	OMPV_VIBCH
RKATVDLGLNADI	115	0.5730	
SANQWLVGATVAY	243	0.9815	
DKFALGAGMNVNF	127	0.9235	OMP1_HAEIN
WGFGWNAGVMYQF	242	0.8520	
NNSRVALGASYNL	349	0.9550	
DRTWYSLGATYKF	390	0.9950	
ENASLELAMAYNF	186	0.9050	S31475
DDTEFVVGIQVEA	398	0.6785	
DRDEITLGASYNF	221	0.9980	OM32_COMAC
KAHQITLGYVHNL	289	0.8660	
NKDASTLGLQAKG	316	0.5995	
SQTGVQVGIRHAF	339	0.9985	
ISDSLNAVAAVAF	68	0.6510	OMPH_PHOS9
TNDELWVGVAGDF	88	0.9410	
NKTTFAVGYTYWS	260	0.6030	
SEFGYVAGMEVTF	314	0.8345	
RNQIWSLGAGYGS	179	0.6225	AFAGBD_9
SQEVFAAGAAYSF	243	0.9700	
SQAVLRVGLRHKF	372	0.9985	
KEFSFKLGGRLQA	54	0.5030	PORP_PSEAE
PGNVVHLGLQYAY	224	0.9705	
EIGAWELFYRYDS	357	0.5450	
SGDGLVMRLQYVF	428	0.8470	
NGESYHVGLNYQN	176	0.9675	OMB2_NEIGO
SQTEVAATAAYRF	264	0.9845	
TYDQVVVGAEYDF	301	0.9990	
VSTASAVVLRHKF	336	0.8755	
RHDDMPVSVRYDS	157	0.7330	OMA1_NEIME
GSDVYYAGLNYKN	215	0.7330	
STTEIAATASYRF	307	0.9735	
SYDQIIAGVDYDF	345	0.9950	
QINAASVGLRHKF	381	0.9830	
AVSSLGLSTIYDF	145	0.6565	OMPC_NEIGO
KTHEASLGMRYRF	258	0.9985	
* *** *******			

Fig. 5. Motif detected in bacterial porins of known structure and related outer membrane proteins. Nineteen bacterial proteins were searched for a single motif (11-column model) with a prior expectation of 100 repeats. The structures of PORI_RHOCA (*R. capsulatus* porin), OMPF_ECOLI (*E. coli* OmpF), and PORI_RHOCA (*R. blastica* porin) are known (see Fig. 6). The alignment is the optimal out of 1,000 independent runs. Proteins are designated by their SwissProt, PIR, or GenBank identifiers.

residues at each position in the alignment with a small number of residue pseudocounts.[4] The target probabilities are given by

$$q_{i,r} = \frac{c_{i,r} + b_r}{c + b} \tag{1}$$

where $c_{i,r}$ is the number of residues of type r at alignment position i (from 1 to the motif width w), b_r is the number of pseudocounts of type r, c is the number of segments in the alignment, and b is the total number of residue pseudocounts. The pseudocounts are distributed among the b_r proportional to the background probabilities (q_r), which are just the amino acid frequencies in the input sequence set. The goal is to identify the most probable motif models by locating those alignments (called optimum alignments) that maximize the ratios of the corresponding target probabilities to the background probabilities (i.e., the likelihood ratios).

More specifically, consider k different motifs of lengths w_1, $w_2, \ldots, w_k$ in a set of S sequences with lengths $\ell_1, \ell_2, \ldots, \ell_S$, so that there are at most

$$N_i = \sum_{j=1}^{S} \max(0, \ell_j - w_i + 1) \tag{2}$$

possible sites for the ith motif. This situation is represented by $k + 1$ motif models $M_0, M_1, M_2, \ldots, M_k$, where M_0 is the null model having target probabilities equal to the background probabilities. Let n_i represent the number of sites that match the ith motif. Although the n_i are initially unknown, given what is known about the biology of the sequences being analyzed and depending on the desired level of stringency (i.e., the amount of similarity shared by the segments in the alignment), a prior expectation e_i for each n_i can be made. (In Bayesian statistics $e_i \div N_i$ corresponds to the prior probability that the ith motif will occur at an arbitrary site in the sequences.) As described below and in the Appendix, the algorithm updates these prior expectations to posterior expectations as it adaptively learns from the data the number of segments corresponding to each motif.

The sampler is initialized by randomly selecting e_i nonoverlapping segments for each motif $M_{i=1\ldots k}$ to create the initial alignments. Then it iteratively performs the following two steps. (1) Select a site (in succeeding iterations, this step is applied to succeeding sites in the sequences); if this site is in one of the alignments, remove it and recalculate the target probabilities for the corresponding model. (2) Sample one of the models (possibly the null model) proportional to the likelihood that the selected site was derived from that model. In doing this, each model is weighted by the posterior probability p_j that an arbitrary site belongs in that model (see Appendix) so that the jth model is sampled proportional to

$$\frac{p_j}{1 - p_j} \prod_{i=1}^{w} \frac{q_{j,i,s_i}}{q_{s_i}} \tag{3}$$

where s_i is the residue observed at the ith position of the segment at the site, $q_{j,i,r}$ is the target frequency for residue r at position i of the jth model, and q_r is the background frequency of residue r.

[4] Pseudocounts arise naturally in the Bayesian approach we have taken and avoid zero probabilities for unobserved residues.

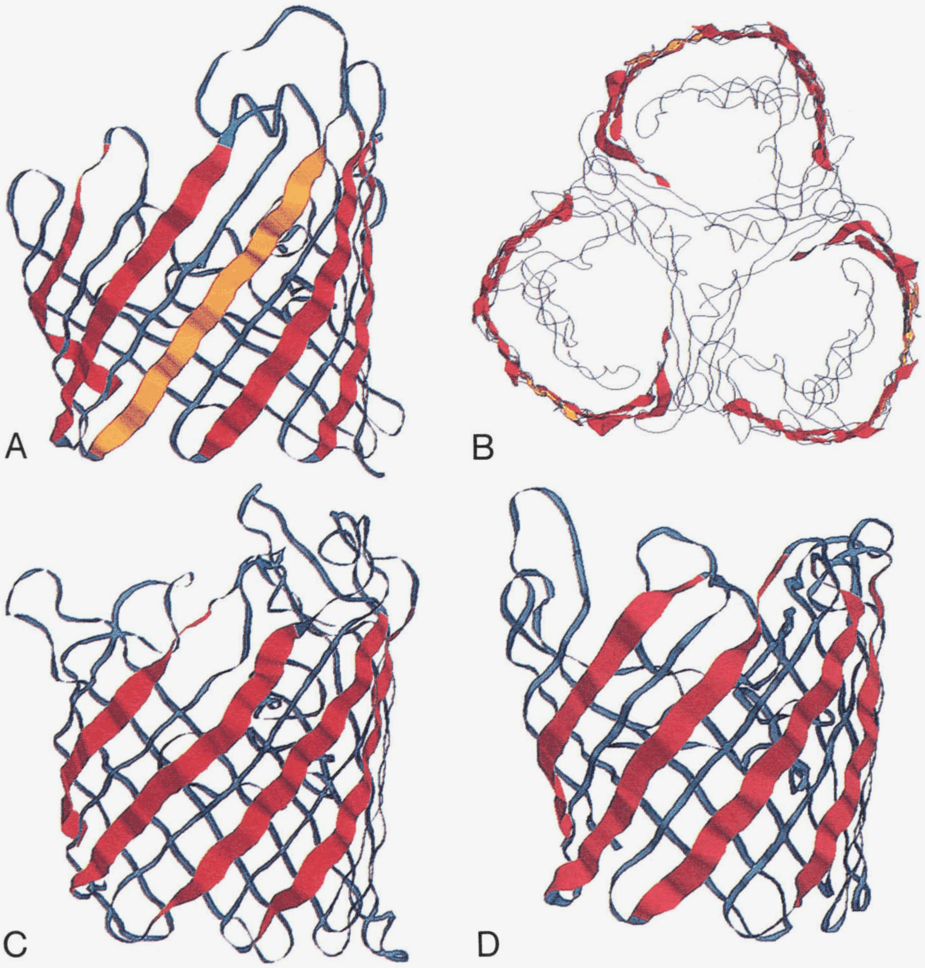

Fig. 6. Locations of conserved repeats within bacterial porins. Tracing of α-carbons are shown as ribbons or strands. Aligned segments from Figure 5 are highlighted in red. **A:** *R. capsulatus* porin (3POR). The segment at site 228 in PORI_RHOCA ("DHKAYGLSVD STF"), although not detected by the sampler, is highlighted in orange because of its location and relatively high probability of matching the motif ($P =$ 0.258) (by default $P \geq 0.50$ is required for detection). **B:** Trimeric *R. capsulatus* porin viewed from above, showing that the conserved repeats occur at the membrane interface. **C:** *E. coli* OmpF porin (1OMF). **D:** *R. blastica* porin (1PRN).

Intuitively, the reason this simple iterative procedure works is that the more accurate the models constructed in step 1, the more accurate are the sites selected for those models in step 2 and vice versa. Consequently, once a few of the correct segments have been selected by chance the model favors the selection of additional correct segments in further iterations, ultimately converging on the optimum alignment(s).

Column sampling

Positions in a polypeptide chain that are important for protein structure or function often tolerate few substitutions. We describe these positions as being information rich because they contribute the most information about the locations of motifs. Well-conserved positions in locally aligned protein sequences, however, are often separated by less conserved (or information poor) positions. For example, the three most informative positions in an alignment of hth regions from DNA-binding proteins are separated by positions containing substantially less information (Fig. 7). Consequently, because the sampler detects subtle patterns by optimizing the information content of the evolving motif model(s), a more nearly optimum motif width may be obtained by using only the most informative positions.

This is accomplished by introducing the notion of fragmentation where only C columns, out of a specified number of contiguous columns $w_{\max} \geq C$, are used ("turned-on") in the residue frequency model. Then, using a column sampling procedure, an initially contiguous C-column model is fragmented by iteratively applying the following two steps. (1) Select an on-column either at random or proportional to how information poor it is and turn it off. (2) Sample one of the $w_{\max} - C + 1$ off-columns proportional to how information rich it is and turn it on. Specifically, the probability of sampling the ith column into the model is proportional to

$$\prod_{r} \left[\frac{\Gamma(c_{i,r} + b_r)}{\prod_{r} q_r^{c_{i,r}}} \right] \tag{4}$$

where $\Gamma(\ )$ is the gamma function (the theoretical basis for Equation 4 is given in Liu et al. [1995]). Thus, after each iteration, unless the same column happens to be chosen in both steps, a column of the evolving model will have moved. In order to avoid biasing the model toward longer widths, however, these column move operations need to be weighted, as is described in the Appendix. Alternating between column sampling and motif sampling can increase the likelihood of converging on the optimum alignment because as the motif sampler improves the evolving alignment this increases the column sampler's ability to locate the most informative columns and vice versa.

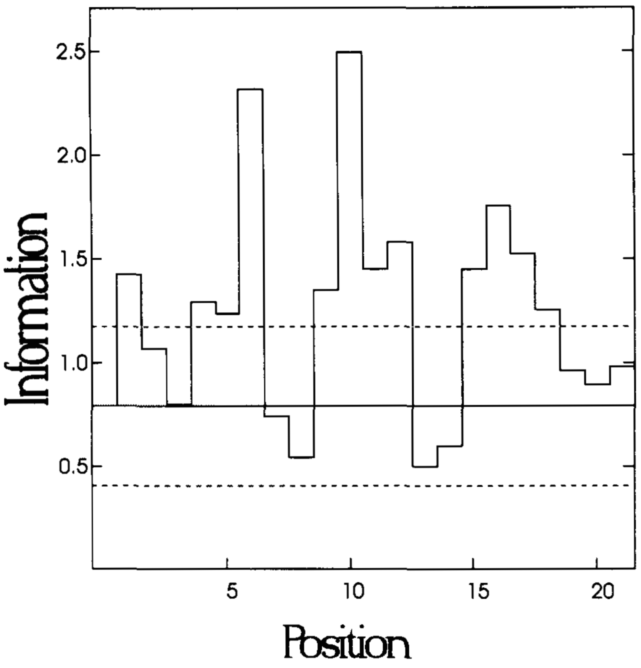

Fig. 7. Position dependence of information content within the hth motif. Estimated information (in bits) corresponding to alignment positions for the 30 hth regions described by Lawrence et al. (1993). The expected information content under the null model is 0.785 bits (solid line) with an SD of 0.191 (dashed lines correspond to ±2 SD); expected values were estimated using the average of 10 optimal alignments of randomly shuffled sequences. Note that three high information sites (positions 6, 10, and 16) are separated by low information sites (positions 8 and 13).

Near-optimum sampling

For subtle motifs, there are often many closely related alignments that are near the optimum. Consequently, a motif site may not be present in the single best alignment found, even though it is present in many of the near-optimum alignments. In order to best identify those sites that are most likely to contain the motif, the following procedure (called near-optimum sampling) is used. After one or more independent runs (as specified by the user), the sampler is reinitialized with the sites obtained from the best alignment found (called the starting alignment). Then sampling continues from among the near-optimum alignments for a sufficient number of cycles (e.g., 1,000–2,000). The fraction of times that a particular site is included in a motif model is the predictive probability that the site matches that motif. By default, those sites that are sampled at least 50% of the time are selected for the final alignment (50% is selected as a cutoff because these sites are more likely than not to contain the motif).

During near-optimum sampling, the model's column configuration is fixed in the starting alignment state, the model's prior expectation (e_i) is set equal to the observed number of segments in the starting alignment, and only those sites having a significant chance of matching the starting alignment model are considered. These modifications keep the sampler from wandering away from the ensemble of alignments that are closely related to the starting alignment. In instances where the sampler fails to find an optimum starting alignment, empirical analysis reveals that applying near-optimum sampling consistently improves the (final) alignment as measured by its likelihood ratio.

Wilcoxon signed rank test of significance

Because the sampler will find the best alignment present in the input sequences, even chance "motifs" can look convincing. Therefore, a statistical test is crucial to evaluating such alignments especially when the detected patterns are subtle. Because many parameters are optimized during the sampling procedure (i.e., the target probabilities and column configurations of the motif models, and the number of segments in the alignments), it is difficult to determine statistical significance analytically. Consequently, we have developed a nonparametric test (Liu et al., 1995) that does not require a knowledge of the underlying probability distribution; it is based on the Wilcoxon (1945) signed rank test. This test requires a single control set of shuffled sequences having the same lengths and overall composition as the input sequences. Under the null hypothesis, sites in the final alignment are just as likely to be drawn from the test set as from the control set. The statistical significance of motifs is measured by determining whether an excess of the best sites was drawn from the test set as follows.

The motif sampler is applied to an input set consisting of both the test and the control sequences. Then the m segments in the final alignment are ranked by decreasing near-optimum sampling frequency (for example, the segments sampled least and most frequently have rank 1 and m, respectively) and control set ranks are given a negative sign. Under the null hypothesis, the mean rank is expected to be near zero, but if the test sequences contain a statistically significant motif, then a significantly large positive mean rank will be found (as determined using a normal approximation or an exact table derived by Wilcoxon [1945]). This test can also be used with the Gibbs site sampler by pairing each test sequence with a control sequence and sampling from among the available sites in both sequences.

Searching a database for matches to motifs

Once a motif or a group of motifs is found, it is often informative to search through a database for additional matching sequences. Given an alignment corresponding to a specific motif, a profile can be constructed by the method of Gribskov et al. (1987, 1990), using linear weighting and a blosum62 scoring matrix (Henikoff & Henikoff, 1992). To determine the probability p_s of obtaining an (ungapped) profile score of at least s for a specific sequence segment, we use the method of Staden (1989), which sums the probabilities associated with every possible segment having a score greater than or equal to s. That is,

$$p_s = \sum_{r_1} \cdots \sum_{r_w} \left(s \leq \sum_{i=1}^{w} S_{i,r_i} \right) \prod_{i=1}^{w} q_{r_i} \qquad (5)$$

where r_i is the residue at position i in the segment, w is the length of the segment, q_r is the frequency of residue r in the database, $S_{i,r}$ is the score corresponding to position i and residue r in the profile, and $(s \leq \sum_{i=1}^{w} S_{i,r_i})$ is a boolean statement having value 1 if true and value 0 if false. (The Staden method computes p_s using an efficient recursive algorithm that requires integer profile scores.) Because p_s is determined using the highest scoring segment in a given sequence, however, a simple Bonferroni adjustment for multiple hypotheses (Weisberg, 1985) is applied by multiplying p_s by the number of segments examined (i.e., by $\ell - w + 1$, where ℓ is the sequence length).

If the motif is internally repeated, a more general form of this method can be used to estimate the probability of finding at least R repeats with scores of at least s. This is done by modeling each hit as a Poisson random event with an expectation of

$$\lambda = (\ell - w + 1) \times p_s. \tag{6}$$

The probability of finding at least R such repeats in a sequence is then given by

$$P = \sum_{x=R}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!}. \tag{7}$$

In order to better determine the optimum number of repeats, Equation 7 is applied to all R over some prespecified range from $R_{\min}$ to $R_{\max}$ (for example, from 2 to 10 repeats). To adjust for multiple R, this probability is multiplied by

$$\begin{aligned} &2^{R-R_{\min}+1} \quad \text{if } R \neq R_{\max} \\ \text{or } &2^{R-R_{\min}} \quad \text{if } R = R_{\max} \end{aligned} \tag{8}$$

according to the weighting scheme of Neuwald and Green (1994). Note, however, that in this case, some caution is needed in interpreting significant hits involving highly similar repeats because the probability is based not only on the distinguishing features of the motif but also on the number of repeats.

For searches involving multiple motifs occurring in a specific order, the individual motifs are linked into a single profile, and, for each sequence in the database, a linear arrangement of nonoverlapping segments with a maximum score s is found. Calculation of p_s is then similar to the single motif case, except that p_s is adjusted for the number of ways that the segments can be selected, that is, by multiplying by

$$\binom{k + \ell - \sum_{i=1}^{k} w_i}{k} \tag{9}$$

where k is the number of motifs.

For each of these cases, a second Bonferroni adjustment is made for the number of sequences in the database. (Note that all of the probability adjustments are conservative.)

Implementation

The motif sampling and database motif search methods described above were implemented as the C language programs GIBBS and SCAN, respectively. (The original Gibbs site sampler is retained as a GIBBS program option.) The default setting for $w_{\max}$ in the GIBBS program is 5 times the number of columns C in the motif model; final alignments are based on 2,000 near-optimum samples. The method of Claverie and States (1993) has been incorporated into the GIBBS program in order to allow optional masking of low complexity regions (Wootton & Federhen, 1993). A C language program PURGE implements a method to reduce sequence redundancy in protein sets. PURGE first computes the maximal segment pair score for every pair of sequences in the input set using a blosum62 scoring matrix (MSP scores are defined by Altschul et al. [1990]). Then it iteratively removes the sequence in the set having the most MSP scores at or above a specified cutoff score s, until only sequences having pairwise MSP scores less than s remain. The source code for these programs is available via anonymous ftp at ncbi.nlm.nih.gov.

Acknowledgments

We thank David Landsman and Jean-Michel Claverie for critical reading of the manuscript and Stephen Altschul and John Spouge for helpful discussions.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403-410.

Bairoch A, Boeckmann B. 1992. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res 20*:2019-2022.

Baldi P, Chauvin Y, McClure M, Hunkapiller T. 1994. Hidden Markov models of biological primary sequence information. *Proc Natl Acad Sci USA 91*:1059-1063.

Barker WC, George DG, Mewes HW, Pfeiffer F, Tsugita A. 1993. The PIR-international databases. *Nucleic Acids Res 21*:3089-3092.

Bennet PB, Makita N, George AL. 1993. A molecular basis for gating mode transitions in human skeletal muscle Na$^+$ channels. *FEBS Lett 326*: 21-24.

Benson D, Lipman DJ, Ostell J. 1993. GenBank. *Nucleic Acids Res 21*:2963-2965.

Bork P, Holm L, Sander C. 1994. The immunoglobulin fold: Structural classification, sequence patterns and common core. *J Mol Biol 242*:309-320.

Bosch D, Scholten M, Verhagen C, Tommassen J. 1989. The role of the carboxy-terminal membrane-spanning fragment in the biogenesis of *Escherichia coli* K12 outer membrane protein PhoE. *Mol Gen Genet 216*: 144-148.

Brennan RG, Matthews BW. 1989. The helix-turn-helix DNA binding motif. *J Biol Chem 264*:1903-1906.

Cardon LR, Stormo GD. 1992. Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J Mol Biol 223*:159-170.

Chan SC, Wong AK, Chiu DK. 1992. A survey of multiple sequence comparison methods. *Bull Math Biol 54*:563-598.

Claverie JM, States DJ. 1993. Information enhancement methods for large scale sequence analysis. *Comput & Chem 17*:191-201.

Cowan SW. 1993. Bacterial porins: Lessons from three high-resolution structures. *Curr Opin Struct Biol 3*:501-507.

Cowan SW, Schirmer T, Rummel G, Steiert M, Ghosh R, Pauptit RA, Jansonius JN, Rosenbusch JP. 1992. Crystal structures explain functional properties of two *E. coli* porins. *Nature 358*:727-733.

Gallegos MT, Michan C, Ramos JL. 1993. The XylS/AraC family of regulators. *Nucleic Acids Res 21*:807-810.

Gribskov M, Luthy R, Eisenberg D. 1990. Profile analysis. *Methods Enzymol 183*:146-159.

Gribskov M, McLachlan M, Eisenberg D. 1987. Profile analysis: Detection of distantly related proteins. *Proc Natl Acad Sci USA 84*:4355-4358.

Harpaz Y, Chothia C. 1994. Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. *J Mol Biol 238*:528-539.

Henikoff S, Henikoff JG. 1991. Automatic generation of protein blocks for database searching. *Nucleic Acids Res 19*:6565-6572.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA 89*:10915-10919.

Henikoff S, Henikoff JG. 1994. Protein family classification based on searching a database of blocks. *Genomics 19*:97-107.

Hunkapiller T, Hood L. 1986. The growing immunoglobulin gene superfamily. *Nature 323*:15-16.

Jap BK, Walian PJ, Gehring K. 1991. Structural architecture of an outer membrane channel as determined by electron crystallography. *Nature 350*:167-70.

Jeanteur D, Lakey JH, Pattus F. 1991. The bacterial porin superfamily: Sequence alignment and structure prediction. *Mol Microbiol 5*:2153-2164.

Jeanteur D, Lakey JH, Pattus F. 1993. The porin superfamily: Diversity and common features. In: Ghuysen JM, Hakebeck R, eds. *Bacterial cell wall.* Amsterdam: Elsevier. pp 363-380.

Jin S, Sonenshein AL. 1994. Identification of two distinct *Bacillus subtilis* citrate synthase genes. *J Bacteriol 176*:4669-4679.

Jones EY. 1993. The immunoglobulin superfamily. *Curr Opin Struct Biol 3*:846-852.

Kaufmann A, Stierhof YD, Henning U. 1994. New outer membrane-associated protease of *Escherichia coli* K-12. *J Bacteriol 176*:359-367.

Kreusch A, Neubuser A, Schiltz E, Weckesser J, Schulz GE. 1994. Structure of the membrane channel porin from *Rhodopseudomonas blastica* at 2.0 Å resolution. *Protein Sci 3*:58-63.

Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. 1994. Hidden Markov models in computational biology: Applications to protein modeling. *J Mol Biol 235*:1501-1531.

Kuma K, Iwabe N, Miyata T. 1991. The immunoglobulin family. *Curr Opin Struct Biol 1*:384-393.

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. 1993. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science 262*:208-214.

Lawrence CE, Reilly AA. 1990. An expectation maximization algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins Struct Funct Genet 7*:41-51.

Lemke G, Lamar E, Patterson J. 1988. Isolation and analysis of the gene encoding peripheral myelin protein zero. *Neuron 1*:73-83.

Leslie DL, Cox J, Lee M, Titball RW. 1993. Analysis of a cloned *Francisella tularensis* outer membrane protein gene and expression in attenuated *Salmonella typhimurium*. *FEMS Microbiol Lett 111*:331-335.

Liu JS, Neuwald AF, Lawrence CE. 1995. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J Am Statist Assoc.* Forthcoming.

Luthy R, Xenarios I, Bucher P. 1994. Improving the sensitivity of the sequence profile method. *Protein Sci 3*:139-146.

Mackett M, Conway MJ, Arrand JR, Haddad RS, Hutt-Fletcher LM. 1990. Characterization and expression of a glycoprotein encoded by the Epstein-Barr virus *Bam*HI I fragment. *J Virol 64*:2545-2552.

Manella CA, Wenwald AF, Lawrence CE. 1996. Identification of likely transmembrane β-strand regions in sequences of mitochondrial pore proteins using the Gibbs sampler. *J Bioenerg Biomembr.* Forthcoming.

Morona R, Klose M, Henning U. 1984. *Escherichia coli* K-12 outer membrane protein (OmpA) as a bacteriophage receptor: Analysis of mutant genes expressing altered proteins. *J Bacteriol 159*:570-578.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol 48*:443-453.

Neuwald AF, Green PP. 1994. Detecting patterns in protein sequences. *J Mol Biol 239*:698-712.

Nikaido H. 1992. Porins and specific channels of bacterial outer membranes. *Mol Microbiol 6*:435-442.

Nikaido H. 1994. Porins and specific diffusion channels in bacterial outer membranes. *J Biol Chem 269*:3905-3908.

Pabo CO, Sauer RT. 1992. Transcription factors: Structural families and principles of DNA recognition. *Annu Rev Biochem 61*:1053-1095.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA 85*:2444-2448.

Pohlner J, Halter R, Beyreuther K, Meyer TF. 1987. Gene structure and extracellular secretion of *Neisseria gonorrhoeae* IgA protease. *Nature 325*:458-462.

Schirmer T, Cowan SW. 1993. Prediction of membrane-spanning β-strands and its application to maltoporin. *Protein Sci 2*:1361-1363.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J Mol Biol 147*:195-197.

Staden R. 1989. Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci 5*:89-96.

Stout V, Torres-Cabassa A, Maurizi MR, Gutnick D, Gottesman S. 1991. RcsA, an unstable positive regulator of capsular polysaccharide synthesis. *J Bacteriol 173*:1738-1747.

Struyve M, Moons M, Tommassen J. 1991. Carboxy-terminal phenylalanine is essential for the correct assembly of a bacterial outer membrane protein. *J Mol Biol 218*:141-148.

Treisman J, Harris E, Wilson D, Desplan C. 1992. The homeodomain: A new face for the helix-turn-helix? *Bioessays 14*:145-150.

Viale AM, Kobayashi H, Akazawa T, Henikoff S. 1991. rbcR, a gene coding for a member of the LysR family of transcriptional regulators, is located upstream of the expressed set of ribulose 1,5-bisphosphate carboxylase/oxygenase genes. *J Bacteriol 173*:5224-5229.

Vogel H, Jahnig F. 1986. Models for the structure of outer-membrane proteins of *Escherichia coli* derived from Raman spectroscopy and prediction methods. *J Mol Biol 190*:191-199.

Weickert MJ, Adhya S. 1992. A family of bacterial regulators homologous to Gal and Lac repressors. *J Biol Chem 267*:15869-15874.

Weisberg S. 1985. *Applied linear regression.* New York: Wiley. p 116.

Weiss MS, Wacker T, Weckesser J, Welte W, Schulz GE. 1990. The three-dimensional structure of porin from *Rhodobacter capsulatus* at 3 Å resolution. *FEBS Lett 267*:268-272.

Wilcoxon F. 1945. Individual comparisons by ranking methods. *Biometrics 1*:80-83.

Williams AF, Barclay AN. 1988. The immunoglobulin superfamily—Domains for cell surface recognition. *Annu Rev Immunol 6*:381-405.

Wootton JC, Federhen S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Comput & Chem 17*:149-163.

Appendix

Prior and posterior motif sampling probabilities

The prior probability of sampling motif model M_i for an arbitrary site is given by

$$p_i = \frac{a_i}{A_i} = \frac{e_i}{N_i} \tag{A1}$$

and the corresponding posterior probability is given by:

$$p_i = \frac{m_i + a_i}{N_i + A_i} \tag{A2}$$

where m_i and a_i are the number of sites and pseudosites, respectively, that are associated with the model, N_i is the total number of sites, and A_i is the total number of pseudosites. As the sampler cycles through the data, the probability of sampling the model for an arbitrary site gets updated continually based on the observed number of sites in the model as formulated by Equation A2.[5] The parameters a_i and A_i determine how much influence the data have on p_i; when A_i is greater than N_i, the pseudosites a_i will carry more weight than the observed sites, m_i, and when A_i is less than N_i, the converse is true. (In Bayesian statistics the number of pseudosites specifies the degree of belief in the prior expectation.) For convenience, we use a fractional weight W to specify the A_i such that

$$A_i = N_i \frac{W}{1 - W} \quad \text{and} \quad a_i = e_i \frac{W}{1 - W} \quad \text{where} \quad 0 < W < 1. \tag{A3}$$

The default setting for W is 0.8.

Weighted column moves

The number of possible column configurations for a C-column model of width w is given by

$$\binom{w - 2}{C - 2} \tag{A4}$$

where $w \geq C \geq 2$. Note that, given a fixed number of columns, the wider models have a greater number of possible column configurations than do the narrower models; if $C = 10$, for example, then there

[5] If the prior expectation e_i is small, then early updating of the probabilities using Equation A2 may cause the evolving alignment to drop rapidly to only one or two sites. This can happen when the model target probabilities, which are computed using the small number of randomly selected segments in the initial alignment, differ significantly from the background probabilities. In this case, unless at least one of the aligned sites contains a motif, candidate sites rarely get sampled, which causes the posterior probability to drop, which then causes even fewer sites to be sampled, and so on. In order to minimize this effect probabilities are updated only after several initial passes through the sequences.

is only one configuration for $w = 10$, but 1,287 configurations for $w = 15$. Thus, using principles similar to those encountered in statistical thermodynamics, it is more likely that the sampler will choose a column configuration corresponding to a wider model simply because the possible configurations (or states) are more numerous. However, all widths can be sampled with equal probability (on average—assuming a random statistical model) if the likelihood of a specific column move (producing a change in width Δw) is multiplied by

$$weight = \frac{\binom{w-2}{C-2}}{\binom{w+\Delta w-2}{C-2}} = \frac{(w-2)!\,(w+\Delta w-C)!}{(w-C)!\,(w+\Delta w-2)!}. \qquad \text{(A5)}$$

Note that this weight is greater than one for negative Δw and less than one for positive Δw.