Proteomics 2004, *4,* 1597–1613 DOI 10.1002/pmic.200300749

Fine-tuning the prediction of sequences cleaved by signal peptidase II: A curated set of proven and predicted lipoproteins of *Escherichia coli* K-12

Pedro Gonnet[1], Kenneth E. Rudd[2] and Frédérique Lisacek[3, 4]

[1]Institute of Computational Science, Swiss Federal Institute of Technology, Zürich, Switzerland
[2]Department of Biochemistry & Molecular Biology, University of Miami School of Medicine, Miami, FL, USA
[3]Geneva Bioinformatics (GeneBio), Geneva, Switzerland
[4]Laboratoire Génome et Informatique, Evry, France

A curated set of 81 proven and 44 predicted lipoproteins of *Escherichia coli* K-12 was defined with the combined use of a literature survey, a variety of predictive tools and human expertise. The well-documented Gram-negative proteome of *E. coli* K-12 was chosen to assess how the different approaches complement each other and to ensure a stable definition of a consistent set of lipoproteins. The results of detailed analysis of such proteins at the level of a single proteome are presented, corroborated and rationalized.

Received	10/11/03
Revised	15/4/04
Accepted	21/1/04

Keywords: Bacterial proteome / Curated set / Lipoprotein / Signal prediction

1 Introduction

Sequence annotation at the level of a single proteome is often automated. Fast processing is particularly needed in the case of prokaryotic proteomes given the current intensive effort for sequencing complete genomes. However, speed and quantity are often achieved to the detriment of quality. This issue is well introduced in [1] where high standards for producing reliable bacterial protein annotations are set. Among others, a relevant strategy involves gathering sequences into consistent families while carefully defining similarity criteria. Several independent initiatives dedicated to grouping bacterial protein into families were launched such as TIGRfam [2], HOBACGEN [3] and HAMAPfam [1] for maximum coverage of protein functions or AraC/XylS [4], among others, for specified functions. Grouping criteria do not necessarily reflect a global similarity of amino acid sequences. Some proteins can be functionally equivalent though structurally very diverse, including at the sequence level. Lipoproteins, with their Type II signal peptides, fall into such a category.

Correspondence: Dr. Pedro Gonnet, HRS H8, ETH Zentrum, Institute of Computational Science, Swiss Federal Institute of Technology, CH-8092 Zürich, Switzerland
E-mail: gonnetp@inf.ethz.ch
Fax: +41-1-632-17-03

Abbreviation: SPI, signal peptidase I

A variety of bacterial lipoprotein prediction schemes are currently available. An early PROSITE pattern (accession number: PS000013, http://www.expasy.org/prosite/) based on [5] and [6] and subsequent InterPro family signature (accession number: IPR000437, http://www.ebi.ac.uk/interpro/) were defined. Scanning tools associated with PROSITE [7] and InterPro [8] can be run to detect the patterns characterizing bacterial lipoproteins. More recently, dedicated computer programs were described [9, 10]. In both cases, the methodology relies on a learning phase during which the program is trained with examples to determine and recognize characteristics of lipoproteins, in other words, to extract and validate a pattern or a motif. A critical issue remains the careful selection of a representative set of examples to ensure the validity of the identified characteristics. In the particular case of bacterial sequences, new instances of motifs are often determined as a result of the presence of orthologues in genome-wide comparisons. Consequently, related and conserved bacterial sequences are commonly gathered into a training set, irrespective of further specificities of each of the organisms. However, distinctive features can also be of interest and such an issue is addressed in the present work.

We have deliberately chosen to focus on the most documented Gram-negative proteome of *Escherichia coli* K-12 and study how the selection of different training sets

bears on the quality of recognition. This strategy helped identify the weaknesses and strengths of published methods as well as the reliability of information found in various sources such as SWISS-PROT [11] and DOLOP [12]. A stable and consistent set of lipoproteins was finally defined which enhances annotation in EcoGene [13], an *E. coli* K-12 dedicated database.

2 Materials and methods

2.1 Training data

We focused on the *E. coli* K-12 proteome. According to publications available at the time of our study, less than 30 *E. coli* lipoproteins were experimentally verified. But recently, Dr. Shin-ichi Matsuyama of the University of Tokyo, experimentally proved the existence of additional *E. coli* lipoproteins. Although these results are not yet formally published, a list of 75 verified lipoproteins provided by Dr. Matsuyama was introduced in the recent LipoP publication [10]. Twenty-one proteins of this list belong to the set of 27 validated lipoproteins as annotated in EcoGene (Version 17). A nonredundant set of 81 proven lipoproteins could potentially be defined as a training set though *FtsI* [14] was omitted since only 15% of *FtsI* is a lipoprotein due to an anomalous cleavage site with a charged arginine residue at position −7 relative to the lipidated cysteine residue. Finally, 81 experimentally verified lipoproteins were used for the positive training set. They are listed along with their corresponding citations in Table 1. None of the signal sequences shared high percentage of similarity, therefore no homology reduction was necessary.

Table 1. Eighty-one experimentally verified *E. coli* K-12 lipoproteins

EG Acc	Gene	SP Acc	Len	Protein Description	SP Annotation	DOLOP	References
EG11703	acrA	P31223	397	AcrAB-TolC efflux pump, membrane-fusion lipoprotein	IPR000437, PS00013	Yes	[10, 26]
EG10266	acrE	P24180	385	AcrEF-TolC efflux pump, membrane-fusion lipoprotein	IPR000437, PS00013	Yes	[10, 53]
EG12073	apbE	P33944	351	Lipoprotein involved in alternative pyrimidine biosynthetic in Salmonella	IPR000437, PS00013	No	[10, 28]
EG12474	blc	P39281	177	Outer membrane lipoprotein, stationary phase inducible	IPR000437, PS00013	Yes	[10, 29]
EG13637	borD	P77330	97	Lipoprotein in DLP12 prophage, phage lambda bor gene homolog	IPR000437, PS00013	Yes	[10, 30]
EG13413	csgG	P52103	277	Possible assembly or transport protein for curli, novel lipoprotein	IPR000437, PS00013	Yes	[10, 31]
EG14233	cusC	P77211	457	Silver and copper efflux, outer membrane lipoprotein component	IPR000437, PS00013	Yes	[10]
EG10178	cyoA	P18400	315	Cytochrome *c* oxidase subunit II, lipoprotein	IPR000437, PS00013	Yes	[32]
EG14372	ecnA	P56548	41	Lipoprotein antidote to bacteriolytic lipoprotein entericidin B	IPR000437, PS00013	No	[33]
EG14345	ecnB	P56549	48	Bacteriolytic lipoprotein entericidin B	IPR000437, PS00013	No	[33]
EG13897	emtA	P76009	203	Membrane-bound transglycosylase, lipoprotein involved in murein hydrolysis	non-EcoGene start	No	[34]
EG10341	ftsI	P04286	423	Septal peptidoglycan synthesis, transpeptidase, 15% of FtsI is lipoprotein	No	No	[14]
EG20264	flgH	P75940	232	Flagellar synthesis, basal body L-ring lipoprotein	IPR000437, PS00013	No	[10, 35]
EG13181	hslJ	P52644	140	Heat-inducible lipoprotein involved in novobiocin resistance	IPR000437	No	[10]
EG11293	lolB	P24208	207	OM lipoprotein required for cell growth and localization of lipoproteins	IPR000437, PS00013	Yes	[10, 36]
EG10544	lpp	P02937	78	Murein lipoprotein	IPR000437, PS00013	Yes	[37, 42]
EG12240	mdtE	P37636	385	MdtEF-TolC multidrug resistance efflux transporter, MFP lipoprotein	IPR000437, PS00013	Yes	[10]
EG11504	metQ	P28635	271	L, D-methionine transporter, methionine-binding lipoprotein receptor	IPR000437, PS00013	Yes	[10]

Table 1. Continued

EG Acc	Gene	SP Acc	Len	Protein Description	SP Annotation	DOLOP	References
EG13085	mltA	P46885	365	mbrane-bound lytic transglycosylase MltA, periplasmic OM lipoprotein	IPR000437, PS00013	Yes	[10, 38]
EG12699	mltB	P41052	361	Membrane-bound lytic transglycosylase MltB, periplasmic OM lipoprotein	IPR000437, PS00013	Yes	[10, 39]
EG12986	mltC	P52066	359	Membrane-bound lytic transglycosylase MltC, periplasmic OM lipoprotein	IPR000437, PS00013	Yes	[10]
EG10246	mltD	P23931	452	Membrane-bound lytic transglycosylase MltD, periplasmic OM lipoprotein	IPR000437, PS00013	Yes	[10]
EG10657	nlpA	P04846	272	Lipoprotein in outer membrane vesicles	IPR000437, PS00013	Yes	[10, 40]
EG10658	nlpB	P21167	344	Lipoprotein in outer membrane vesicles	IPR000437, PS00013	No	[10, 41]
EG11133	nlpC	P23898	154	NlpC lipoprotein	IPR000437, PS00013	Yes	[10]
EG12111	nlpD	P33648	379	Lipoprotein possibly involved in cell wall formation, metalloprotease homolog	IPR000437, PS00013	No	[10, 27]
EG12137	nlpE	P40710	236	Outer membrane lipoprotein, activates Cpx response in response to adhesion	IPR000437, PS00013	No	[10, 54]
EG12371	nlpI	P39833	294	Minor lipoprotein, mutation causes osmotic sensitivity and filamentation	IPR000437, PS00013	Yes	[10, 44]
EG10679	osmB	P17873	72	OsmB lipoprotein	IPR000437, PS00013	Yes	[10, 45]
EG10044	osmE	P23933	112	Lipoprotein regulated by growth phase and osmotic pressure	IPR000437, PS00013	Yes	[10]
EG10684	pal	P07176	173	Lipoprotein associated with peptidoglycan	IPR000437, PS00013	Yes	[10, 46, 47]
EG11502	rcsF	P28633	134	Lipoprotein, overexpression increases capsule synthesis	IPR000437	Yes	[10]
EG10854	rlpA	P10100	362	Minor lipoprotein, suppressor of prc	IPR000437, PS00013	Yes	[10, 48]
EG10855	rlpB	P10101	193	Minor lipoprotein	IPR000437, PS00013	No	[10, 48]
EG11890	slp	P37194	188	Outer membrane lipoprotein, stationary phase inducible	IPR000437, PS00013	No	[10, 49]
EG13409	slyB	P55741	155	Novel lipoprotein, Mg(2+)-stimulated	IPR000437, PS00013	Yes	[10, 50]
EG14076	spr	P77685	188	Suppressor of prc mutants at low osmolality, lipoprotein	IPR000437, PS00013	No	[10]
EG14276	vacJ	P76506	251	Surface-exposed lipoprotein, required for intercellular spreading in Shigella	IPR000437, PS00013	Yes	[10]
EG13566	wza	P76388	379	Outer membrane auxillary lipoprotein, capsular polysaccharide translocation	IPR000437, PS00013	Yes	[10, 51]
EG13332	yafT	P77339	261	Novel lipoprotein, function unknown	IPR000437, PS00013	Yes	[10]
EG13253	ybaY	P77717	190	Novel lipoprotein, function unknown	IPR000437, PS00013	Yes	[10]
EG13660	ybfN	P75734	108	Novel lipoprotein, function unknown	IPR000437, PS00013	Yes	[10]
EG14158	ybfP	P75737	164	Novel lipoprotein, function unknown	IPR000437, PS00013	Yes	[10]
EG12875	ybhC	P46130	427	Novel lipoprotein, pectinesterase homolog, function unknown	No	Yes	[10]
EG13685	ybjP	P75818	171	Novel lipoprotein, function unknown	IPR000437, PS00013	Yes	[10]
EG13687	ybjR	P75820	276	Novel lipoprotein, homologous to AmpD, function unknown	IPR000437	Yes	[10]
EG13133	ycaL	P43674	254	Novel lipoprotein, metalloprotease homolog, function unknown	IPR000437	No	[10]
EG13728	yccZ	P75881	379	Novel lipoprotein, Wza paralog, function unknown	IPR000437, PS00013	Yes	[10]
EG13864	ycdR	P75906	672	Polysaccharide deacetylase-like lipoprotein, function unknown	IPR000437, PS00013	Yes	[10]
EG11117	yceB	P09995	186	Novel lipoprotein, function unknown	IPR000437, PS00013	Yes	[10]
EG12689	yceK	P45806	75	Novel lipoprotein, function unknown	IPR000437	Yes	[10]
EG13431	ycfM	P75947	213	Novel lipoprotein, function unknown	IPR000437	Yes	[10]

Table 1. Continued

EG Acc	Gene	SP Acc	Len	Protein Description	SP Annotation	DOLOP	References
EG13911	ycjN	P76042	430	Putative ABC transporter periplasmic binding lipoprotein, function unknown	IPR000437	Yes	[10]
EG13755	ydcL	P76101	222	Novel lipoprotein, function unknown	IPR000437, PS00013	Yes	[10]
EG13794	yddW	P76130	439	Novel lipoprotein, function unknown	IPR000437, PS00013	No	[10]
EG13511	yeaY	P76255	193	Novel lipoprotein, slp paralog, function unknown	IPR000437, PS00013	Yes	[10]
EG14036	yecR	P76308	107	Novel lipoprotein, function unknown	IPR000437, PS00013	Yes	[10]
EG11659	yedD	P31063	137	Novel lipoprotein, function unknown	IPR000437, PS00013	Yes	[10]
EG12004	yehR	P33354	153	Novel lipoprotein, function unknown	IPR000437, PS00013	Yes	[10]
EG14166	yfeY	P76537	191	Novel lipoprotein, function unknown	IPR000437, PS00013	Yes	[10]
EG14204	yfgH	P76572	172	Novel lipoprotein, function unknown	IPR000437	Yes	[10]
EG14208	yfgL	P77774	392	Novel lipoprotein, function unknown	IPR000437	Yes	[10]
EG11152	yfiB	P07021	160	Putative outer membrane lipoprotein, ompA homolog, function unknown	IPR000437, PS00013	Yes	[10]
EG12446	yfiL	P11289	121	Novel lipoprotein, function unknown	IPR000437	Yes	[10]
EG14222	yfiO	P77146	245	Novel lipoprotein, homologous to N. gonorrhoeae ComL, function unknown	IPR000437, PS00013	Yes	[10]
EG13081	ygdI	Q46924	75	Novel lipoprotein, ygdR paralog, function unknown	IPR000437, PS00013	Yes	[10]
EG13076	ygdR	Q46932	72	Novel lipoprotein, ygdI paralog, function unknown	IPR000437, PS00013	No	[10]
EG13048	ygeR	Q46798	251	Novel lipoprotein, metalloprotease homolog, function unknown	IPR000437, PS00013	No	[10]
EG12991	yghG	Q46835	136	Novel lipoprotein, function unknown	IPR000437, PS00013	Yes	[10]
EG12833	yhdV	P45765	73	Novel lipoprotein, function unknown	IPR000437, PS00013	Yes	[10]
EG12907	yhfL	P45538	55	Novel lipoprotein, function unknown	IPR000437	Yes	[10]
EG12271	yiaD	P37665	219	Novel lipoprotein, ompA homolog, function unknown	IPR000437, PS00013	Yes	[10]
EG11860	yiiG	P32151	351	Novel lipoprotein, function unknown	IPR000437	Yes	[10]
EG11924	yjbF	P32687	212	Novel lipoprotein, ymcC paralog, function unknown	IPR000437, PS00013	No	[10]
EG12471	yjeI	P39278	117	Novel lipoprotein, function unknown	IPR000437	No	[10]
EG13731	ymcC	P75884	214	Novel lipoprotein, yjbF paralog, function unknown	IPR000437, PS00013	Yes	[10]
EG14298	ynbE	P76075	61	Novel lipoprotein, function unknown	IPR000437	Yes	[10]
EG13841	ynfC	P76171	236	Novel lipoprotein, function unknown	IPR000437, PS00013	No	[10]
EG14304	yoaF	P76244	84	Novel lipoprotein, function unknown	IPR000437	Yes	[10]
EG13018	yqhH	Q46860	85	Novel lipoprotein, lpp paralog, function unknown	IPR000437, PS00013	Yes	[10]
EG12781	yraP	P45467	191	Novel lipoprotein, osmY paralog, function unknown	IPR000437, PS00013	No	[10]

EcoGene includes a compilation of all *E. coli* proteins whose *N*-terminal sequences were experimentally determined. This verified set currently contains 862 proteins: http://bmb.med.miami.edu/EcoGene/EcoWeb/CESSPages/VerifiedProts.htm. Each protein is associated with the corresponding primary literature citations and the number of amino acids removed post-translationally, if any. Two negative training sets were derived from this verified set: (i) a set of 135 exported proteins cleaved by signal peptidase I (SPI), and (II) a set of 722 proteins that are either not cleaved or cleaved by methionine aminopeptidase that liberates the *N*-terminal methionine residue. The second set was used for training the motif parameters whereas the first was used only for later verifications.

The EcoGene database includes revised predicted start sites for 730 *E. coli* proteins. Numerous proteins needed to be shortened as a result of an original annotation strategy that favored the longest ORF as opposed to the most likely [5]. Most of these corrections have been communicated to SWISS-PROT from EcoGene as part of a

collaborative annotation effort. Incorrect or unlikely annotation of translation start sites can cause many problems in postgenomics research, including the identification of potential lipoproteins.

2.2 Lipoprotein motif

We define a motif as a linear sequence of attributes or tokens. Each token describes one or more characteristics of a single amino acid or subsequence. Different token types were set. They are listed in Table 2. The different characteristics are shown in Table 3. The lipoprotein motif used for training is very similar to the one used in [9] for the detection of lipoproteins in *Bacillus subtilis.* In the PATOSEQ syntax it is written as:

M [p,3:3](0,25) !{R,K} [h~{R,K},15:15](6,20) {} {} {} C ✳

Table 2. Different token types and their interpretation. Each token can be weighted by prefixing with a numerical value. Tokens prefixed with a ! are locked for refinement

A	A fixed amino acid (anchor).
a	A variable amino acid (distance measure used: $Dayhoff_{250}$).
{S = 0.4, T = 0.6}	A frequency vector specifying the frequency of each amino acid. If no residue is specified, the natural frequencies of each amino acid are used. If residues are listed with no specified frequency, the relative natural frequency is assumed. If this token is preceded by a tilde (~), values are inverted (exclusive vector).
[x, 10:1]	A sequence of length 10 with variance 1 and the characteristics x. A range (min, max) can optional be appendend.
–	Any single amino acid
*	Any sequence of amino acids

Table 3. The variety of characteristics for sequence tokens. Each characteristic can be weighted by prefixing with a numerical value

p, n, u	positive, negative or no charge
o, y	hydrophobic or hydrophilic
l, s	large or small (volume)
a, b, i	amphipatic alpha helix, beta sheet or volume-helix
{A, G}	frequency vector
*	no characteristics

which should be read as an initial methionine residue (M), followed by a positively charged region of length 0 to 25 ([p,3:3](0,25)), followed by either an arginine or lysine residue (!{R,K}[1]), followed by a hydrophobic region of length 6 to 20 residues characterized by a frequency vector initially not containing positively charged residues ([h~{R,K}, 15:15](6,20)[2]), followed by three residues characterized by frequency vectors ({} {} {}), followed by a fixed cysteine residue (C) which is the lipid binding site, followed by anything (✳).

2.3 Motif scoring

Once a motif is aligned to an amino acid sequence, a corresponding score is calculated as the sum of the partial scores for each aligned token:

$$score(m) = \sum_{i} score_i(t_i) \quad (1)$$

where t_i is the i^{th} token in the motif m and $score_i$ its scoring function. For subsequences containing more than one characteristic, the partial score for each characteristic is summed. To ensure the consistency of the alignment, partial scores must all be in the same unit of measurement, independently of the characteristic being scored. Consequently, the partial score is defined as the relative log-probability of the aligned subsequence fitting the characteristics associated with the given token as opposed to a random match:

$$score(t_i) = \log\left(P^+(t_i)\right) \quad (2)$$

The total score can then be interpreted as the log-probability of the entire aminoacid sequence matching the entire motif:

$$e^{score(m)} = \prod_{i} P^+(t_i) \quad (3)$$

Since some features in a motif can be more important than others, they are optionally weighted. These weights are multiplied with the partial log-probabilities:

$$score(t) = \sum_{i} w_i \log\left(P^+(t_i)\right) \quad (4)$$

$$e^{score(m)} = \prod_{i} P^+(t_i)^{w_i} \quad (5)$$

where w_i is the assigned weight of the i^{th} token. Two tokens t_1 and t_2 with weights w_1 and w_1 are interpreted as the characteristics in t_1 occur $\frac{w_1}{w_2}$ times more often than those in t_2.

The functions for the different $P^+(t_i)$ are defined separately for each token type and each characteristic in a token. For the token types any (✳) and space (_), this probability is

[1] The ! modifier locks this token so that it is not modified during training, effectively forcing the motif to match either R or K.
[2] ~ modifier inverts a given frequency vector.

always 1. For fixed amino acids (A), the probability is 1 in the case of a match and 0 otherwise.

For frequency vector tokens, $P^+(t_i)$ is given by the relative probability of an amino acid a matching the frequency vector f_v compared with natural occurrence f_n:

$$P_f^+(a) = \frac{f_v(a)}{f_v(a) + f_n(a)} \tag{6}$$

For frequency vectors within sequence tokens, the average relative probability is used:

$$P_f^+(s) = \frac{\prod_i f_v(s_i)}{\prod_i f_v(s_i) + \prod_i f_n(s_i)} \tag{7}$$

where s_i is the i^{th} amino acid in the subsequence s. To allow for flexibility (*i.e.*, during refinement), a noise factor ϵ can be added to the relative frequency:

$$P_f^+(a) = \frac{f_v(a) + \epsilon f_n(a)}{f_v(a) + f_n(a)} \tag{8}$$

which also avoids "trapping" the total score at 0 and hindering refinement, as seen later.

The length of a subsequence is similarly scored. The length of a given subsequence is assumed to be Poisson distributed (insertions and deletions being discrete cumulative events). If in the random case the subsequence length is evenly distributed over an interval (a, b), then P_l^+ is defined as:

$$P_l^+(s) = \frac{\pi_\mu(|s|)}{\pi_\mu(|s|) + (b - a + 1)^{-1}} \tag{9}$$

where $\pi_\mu(x)$ is the Poisson probability of x with $\lambda = \mu$. Values for a and b are set to 1 and 100, unless specified otherwise.

The scoring of the other tokens (*i.e.*, charge, hydrophobicity, volume, *etc.*) is somewhat more complicated, since they correspond to a more abstract concept of high or low as opposed to fixed values of these characteristics. Charge, hydrophobicity and volume are estimated from the normal distribution of their expected value, *e.g.*, for charge:

$$\mu_{ch} = \sum_a f_r(a) ch(a) \tag{10}$$

$$\sigma_{ch}^2 = \sum_a f_r(a)(\mu_{ch} - ch(a))^2 \tag{11}$$

where $ch(a)$ is the net charge of the amino acid a. The same scoring scheme is applied for amino acid volume and hydrophobicity indices, and in fact, any index that models a physicochemical characteristic.

Given the average charge of a subsequence s, P_p^+ (positive charge) is then calculated as the inverse probability of the charge being greater in the random case:

$$P_p^+(s) = CDF(\mathrm{N}(\mu_{ch}, \sigma_{ch}), ch(s)) \tag{12}$$

where N(μ, σ) is the normal (Gaussian) distribution over μ and σ and $ch(s)$ is the average charge *per* residue of the subsequence s. This scoring method can be applied to any value for which the distribution can be derived analytically or experimentally (for more examples see [9]).

Finally, the optimal alignment of any motif with an amino acid sequence is achieved using dynamic programming (Fig. 1). This alignment scoring technique significantly differs from scoring with regular expressions and other grammars, or with weight matrices. In our method, a protein sequence is compared to a description of a protein sequence, so that the score is the maximized probability of the sequence fitting the description. In other words, we are not evaluating some abstract numerical score, but the ratio of the probability of the given sequence matching the motif over the probability of a random sequence matching the motif. The resulting score is not binary, as in the case of grammars and regular expressions. Moreover, it is statistically and biologically interpretable, unlike weight matrix scores.

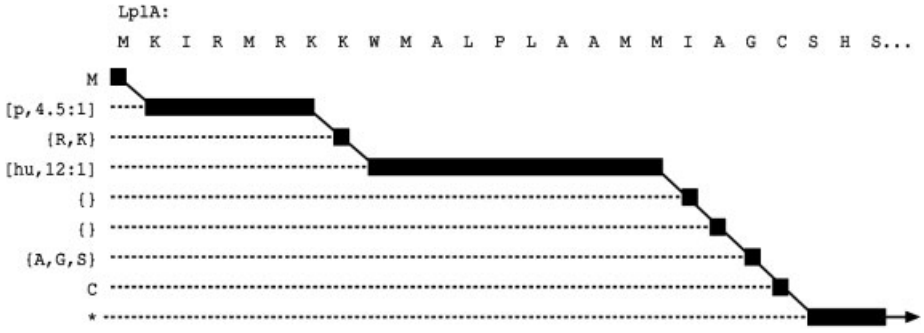

Figure 1. Motif Alignment: the motif is aligned to the sequence using dynamic programming much in the same way as pairwise sequence alignment.

2.4 Motif refinement

Given a set of known (or expected) positive and negative examples and a motif, a classification score is calculated to assess how well the given motif discriminates between the two sets. The motif can then be parameterized to maximize this score. Abundant literature is describing widely used scoring methods (Mathews correlation coefficient [16], Fisher linear discriminant [17], LinearClassify [18], Precision/Recall [19] to quote the most popular). As pointed out in [20], these approaches are almost all directly based on the effective number of misclassified sequences. In the present work, the scoring scheme for classification does not depend on this quantity.

Scores of sequences in a positive or negative dataset can be plotted in a distribution. Such score distributions can be modelled as beta-distributions. The expected percentage of misclassified sequences can be calculated as the overlap between both distributions relative to a cut-off value *c*:

$$disc(c) = \frac{|S^-|}{|S^-|+|S^+|} CDF(\beta(\mu^-, \sigma^-), c) + \frac{|S^+|}{|S^-|+|S^+|}(1 - CDF(\beta(\mu^+, \sigma^+), c)) \quad (13)$$

where S^+ and S^- are the positive and negative sets and $\beta(\mu^+, \sigma^+)$ and $\beta(\mu^-, \sigma^-)$ their beta-distributions. The cut-off value *c* is chosen such as to maximize this score.

This approach is justified since sequences used for training are mere samples of an exhaustive set (*i.e.* an entire proteome). The classification score corresponds to the expected extent of the overlap. The method relies more on the average characteristics of our sequences and outliers are effectively treated as such (*i.e.* exceptions are tolerated). This will later avoid overfitting of the motif to the training set. The classification score is maximized over the motif parameters (frequency vectors, lengths and weights), which is nontrivial. Indeed, any change in the motif can affect the partial scores hence the alignment outcome, leading to discontinuities in the scoring surface.

Refinement is based on a heuristics presented in [9]. The frequency vector values and subsequence lengths are adjusted according to values observed in aligned positive examples. Weights are adjusted according to their relative discriminative power or through a least-square minimization of the overlapping distributions (approximate).

Optimization is iterative. Each parameter change is followed by sequence realignment and the change is kept only if the classification score improves. The motif is initially trained as detailed in [9] using the verified positive and negative training sets described above. The validity of this prediction is then tested using a *k*-fold validation test over the positive training sequences.

The refined motif is then applied to all sequences of the EcoProt (translated EcoGene) database (Version 17). This initial prediction is used for bootstrapping over the EcoProt set of sequences. Bootstrapping involves using the results of a prediction to re-refine the initial motif. It is followed by a new prediction over the same data set based on the re-refined motif. Practically, false positives and false negatives are considered respectively as positives and negatives until no further refinement is possible. The procedure is illustrated in Fig. 2. Bootstrapping is not used to optimize the motif itself, but to optimize the partition of the proteome given an initial motif. Intuitively speaking, such a partition can and should be assumed since the cell is likely to distinguish

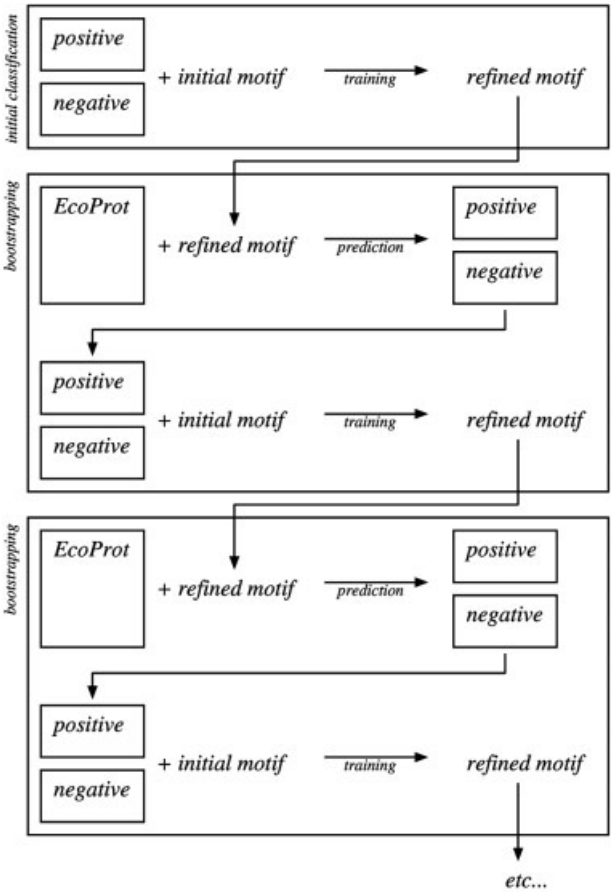

Figure 2. The different steps in motif refinement. Starting from an initial motif and an initial training set, the motif is refined. The refined motif is then used to partition a proteome into positive and negative predictions which are in turn used to re-refine the initial motif and so on, until the predictions remain stable.

between lipoproteins and nonlipoproteins. Moreover, the partition is not likely to degenerate and shift away from a lipoprotein/nonlipoprotein classification, given that only the parameterization of the initial motif is optimized during refinement. As the results show, this is effectively the case.

3 Results

3.1 Motif refinement

The motif was trained as described in Section 2.4. Both sets are perfectly distinguished with the refined motif. The lowest scoring positive is 30.139% and the highest scoring negative, 3.339%. The total classification score over the training data is 100%. In other words, all training sequences are discriminated correctly, and the distributions of scores do not discernibly overlap.

For the *k*-fold cross-validation, the set of known positive sequences was divided into 10 groups of eight sequences and the motif was retrained using all combinations of nine of these 10 groups and tested against the remaining one. The groups were created deterministically by sorting and splitting the data according to EcoGene IDs. This approach is equivalent to a random partition of the data set and was chosen to ensure unbiasedness and reproducibility.

Of the resulting predictions, six positive sequences (7.5%) are missed: *nlpE*, *rlpA*, *rlpB*, *mltC*, *apbE* and *yddW*. All of these sequences except for *rlpA* have singular residues in the lipobox and *RlpA* is the only sequence containing a tryptophan residue in the transmembrane region. These weak characteristics are difficult to predict if they are not represented in the training set. The refined motif was then applied to all sequences of the EcoProt (translated EcoGene) database (Version 17), yielding 120 predicted lipoproteins and 1457 nonlipoproteins. The remaining sequences could not be aligned due to the missing required methionine and cysteine residues and are therefore also considered nonlipoproteins, although they could no longer be used for this analysis.

After a first bootstrapping run, the initial motif was modified to be more specific in the lipobox:

{I,L,M,F,T,V} {} {A,G,S} C ✳

for faster convergence. Since the content of the frequency vectors is also subject to refinement, the exibility of the lipoprotein signal motif is not altered. Bootstrapping converged after two iterations towards 118 predicted lipoproteins with a classification score of 99.98%. *nanE* and *yjiK* were reclassified as nonlipoproteins. The lowest scoring positive sequence is *yddW* with 36.573% and the highest scoring negative is *nanE* with 31.886%. The distribution of the scores can be seen in Fig. 3. The complete set of predicted lipoproteins and their scores are shown in Table 4.

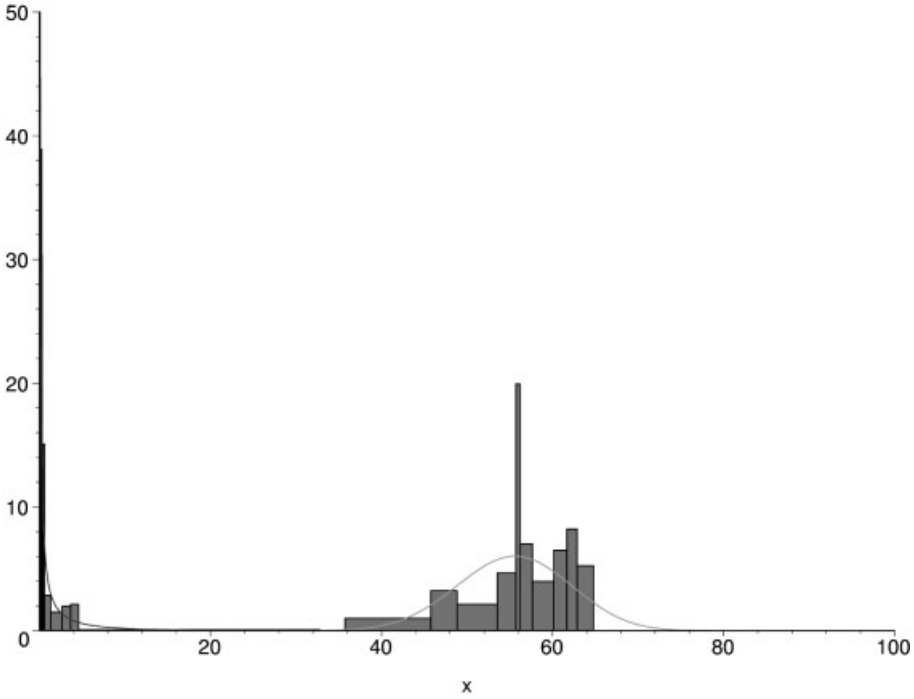

Figure 3. Distribution of the scores of the positive and negative predictions with the fitted β-distributions.

Table 4. The 118 predicted lipoprotein signals found in the EcoGene *E. coli* protein sequences using PATOSEQ

Name	Score	Aligned sequence
mltB	65.10412	M FKR R YVTLLPLFVL L A A C SSK …
mdtE	64.84520	M NRRR K LLIPLLFCGAM L T A C DDK …
yajG	64.76101	M FK K ILFPLVALFM L A G C AKP …
ybjR	64.49154	M R R FFWLVAAALL L A G C AGE …
ecnA	64.17996	M MK R LIVLVLLASTL L T G C NTA …
yafT	64.09877	M NSK K LCCICVLFSL L A G C ASE …
yfbK	64.01340	M RN K NIIMLLMSSLI L S G C GPQ …
yhdV	63.77605	M K R LIPVALLTAL L A G C AHD …
yfgL	63.63713	M QLR K LLLPGLLSVTL L S G C SLF …
yfgH	63.58306	M MKFK K CLLPVAMLASFT L A G C QSN …
yfiO	63.44926	M TRM K YLVAAATLSLF L A G C SGS …
yggG	63.36798	M KI R ALLVAMSVATV L T G C QNM …
yfhM	63.35255	M KKL R VAACMLMLA L A G C DNN …
ybjP	62.65524	M RYS K LTMLIPCALL L S A C TTV …
ycfM	62.44483	M TKMS R YALITALAMF L A G C VGQ …
slyB	62.34039	M IK R VLVVSMVGLS L V G C VNN …
mltA	62.32953	M KGRWV K YLLMGTVVAM L A A C SSK …
ycfL	62.17904	M R K GCFGLVSLVLLL L V G C RSH …
ycaL	62.12220	M KNT K LLLAIATSAAL L T G C QNT …
lpp	62.09211	M KAT K LVLGAVILGSTL L A G C SSN …
ynfC	62.03280	M KY K LLPCLLAIF L T G C DRT …
ymbA	61.93177	M K K WLVTIAALW L A G C SSG …
yifL	61.87079	M KNVF K ALTVLLTLFS L T G C GLK …
yfeY	61.69214	M KSL R LMLCAMPLM L T G C STM …
yccZ	61.39632	M KKNIF K FSVLTLAVLS L T A C TLV …
yedD	61.30897	M K K LAIAGALLL L A G C AEV …
rzoR	61.23163	M RKL K MMLCVMMLPLV V V G C TSK …
rzoD	61.23163	M RKL K MMLCVMMLPLV V V G C TSK …
ycdR	60.95369	M LRNGN K YLLMLVSIIM L T A C ISQ …
ybfN	60.77116	M K K LILIAIMASG L V A C AQS …
ybfN	60.68041	M KR K TLPLLALVATSLF L S A C DDR …
vacJ	60.55754	M KL R LSALALGTTL L V G C ASS …
hslJ	60.49410	M K K VAAFVALSLL M A G C VSN …
yhfL	60.47031	M NKFI K VALVGAVLAT L T A C TGH …
dcrB	60.20883	M RNLV K YVGIGLLVMG L A A C DDK …
yegR	60.02550	M K K IAAISLISIFI M S G C AVH …
nlpI	59.93968	M KPFL R WCFVATALT L A G C SNT …
borD	59.80778	M K K MLLATALALL I T G C AQQ …
yehR	59.78834	M KAFN K LFSLVVASVLVFS L A G C GDK …
wza	59.65025	M MKSKM K LMPLLVSVTL I S G C TVL …
yghJ	59.60455	M NKKFKYK K SLLAAILSATL L A G C DGG …
emtA	59.38242	M KL R WFAFLIVL L A G C SSK …
ycjN	58.98437	M IKS K IVLLSALVSCAL I S G C KEE …
ygdI	58.50586	M K K TAAIISACMLTFA L S A C SGS …
slp	58.37951	M NXT K GALILSLSFL L A A C SSE …
cyoA	58.23754	M RLRKYN K SLGSLSLFAGTVL L S G C NSA …
yfiL	58.10022	M MK K FIAPLLALL V S G C QID …
cusC	57.70727	M SPC K LLPFCVALA L T G C SLA …
yafY	57.57113	M KR K TLPLLALVATTLF L I A C DDR …
yiaD	57.31151	M KK R VYLIAAVVSGALA V S G C TTN …
yoaF	57.18330	M K IISFVLPCLLV L A G C STP …
ydcL	57.11211	M RTTSFA K VAALCGLLA L S G C ASK …
csgG	57.10907	M Q R LFLLVAVML L S G C LTA …
mltD	56.92914	M KA K AILLASVL L V G C QST …
yfiB	56.72722	M I K HLVAPLVFTSLI L T G C QSP …
nlpE	56.70599	M VK K AIVTAMAVISLFT L M G C NNR …
nlpC	56.68813	M R FCLILITALL L A G C SHH …
yghG	56.66591	M SIKQMPG R VLISLLLSVTGL L S G C ASH …
nlpA	56.60377	M KLTTHHL R TGAALLLAGIL L A G C DQS …

Name	Score	Aligned sequence
yjbF	56.56772	M K R PALILICLL L Q A C SAT …
osmB	56.47267	M FVTSK K MTAAVLAITLAMS L S A C SNW …
yceK	56.43614	M R LIVVSIMVTL L S G C GSI …
ymcC	56.40653	M R PLILSIFALF L A G C THS …
rcsF	56.36003	M R ALPICLVALM L S G C SML …
metQ	56.33563	M AFKF R TFAAVGALIGSLA L V C C CQD …
fecD	56.24603	M K IALVIFITLA L A G C ALL …
yeaY	56.22610	M AVQKNVI K GILAGTFALM L S G C VTV …
yecR	56.15763	M R LLILTLSLIT L A G C TVT …
yajI	56.12183	M NTNVF R LLLLGSLFS L S A C VQQ …
ygdR	56.08548	M K K WAVIISAVGLAFA V S G C SSD …
mltC	56.07822	M K K YLALALIAPL L I S C STT …
ygeR	55.96281	M SAGRLNK K SLGIVMLLSVGLL L A G C SGS …
yceB	55.86755	M N K FLFAAALIVSGL L V G C NQL …
yjeI	55.82614	M HV K YLAGIVGAALL M A G C SSS …
ybhC	55.80707	M NTFSVS R LALALAFGVT L T A C SST …
yidX	55.46699	M KLNFKGFF K AAGLFPLALM L S G C ISY …
pal	55.39019	M QLNKVL K GLMIALPVMA I A A C SSN …
ydbJ	55.32699	M R AAFWVGCAALL L S A C SSE …
ecnB	55.22635	M VK K TIAAIFSVLVLSTV L T A C NTT …
ybfP	55.15352	M KTN R SLVVIVSLITATLL L T A C AQP …
yfiM	55.05734	M R ILFVCSLLL L S G C SHM …
mdtP	54.84027	M INRQLS R LLLCSILGSTTL I S G C ALV …
smpA	54.76335	M RC K TLTAAAAVLLML T A G C STL …
yiaF	54.74585	M ATGKSCS R WFAPLAALLMVVS L S G C FDK …
ynBE	54.70140	M K ILLAALTSSFM L V G C TPR …
rlpA	53.21251	M R K QWLGICIAAGM L A A C TSD …
acrE	52.72669	M TKHA R FFLLPSFILISAAL I A G C NDK …
fepG	52.51321	M IYVSR R LLITCLLL V S A C VVA …
yaeF	52.40728	M DKPKAYC R LFLPSFLL L S A C TVD …
ybaY	52.12813	M K LVHMASGLAVAIA L A A C ADK …
yliB	51.92793	M ARAVH R SGLVALGIATAL M A S C AFA …
yaiW	51.48457	M S R VNPLSSLSLLAVLV L S A C SSQ …
yebF	51.48306	M EKNMKK R GAFLGLLL V S A C ASV …
yiiG	50.81708	M K R NLLSSAIIVAIMSLG L T G C DDK …
osmE	50.61636	M N K NMAGILSAAAVLTM L A G C TAY …
ydjY	50.50179	M LQHYSVSWK K GLAALCLLAVAG L S G C DQQ …
yqhH	49.55094	M K TIFTVGAVVLATCL L S G C VNE …
fliL	49.10648	M TDYAISKKSK R SLWIPILVFIT L A A C ASA …
nlpD	49.09852	M SAGSPKFTVR R IAALSLVSLW L A G C SDT …
yraP	48.60274	M K ALSPIAVLISALL L Q G C VAA …
yjfO	48.43271	M VSRKRNSVIY R FASLLLVLM L S A C SAL …
lolB	48.15941	M PLPDFRLI R LLPLAALV L T A C SVT …
blc	47.84833	M R LLPLVAAATAAFL V V A C SSP …
ypdI	47.50745	M K VNLILFSLFLLVS I M A C NVF …
nlpB	47.26330	M AYSVQKSRLA K VAGVSLVLL L A A C SSD …
apbE	47.21592	M EISFT R VALLAAALF F V G C DQK …
rlpB	46.96210	M R YLATLLLSLAVLI T A G C GWH …
acrA	45.26031	M NKN R GFTPLAVVLMLSGSLA L T G C DDK …
arnT	44.25810	M KSV R YLIGLFA F I A C YYL …
spr	44.23807	M VKSQPILRYIL R GIPAIAVAVL L S A C SAN …
flgH	44.14161	M Q K NAAHTYAISSLLVLS L T G C AWI …
yqiG	43.01175	M I K NIYCSLSVI I I G C ASA …
rtn	42.89918	M FIRAPNFGR K LLLTCIVAGVMIAI L V S C LQF …
ampC	42.79648	M F K TTLCALLI T A S C STF …
yraM	41.24680	M VPSTFSRLKAA R CLPVVLAALI F A G C GTH …
yecT	40.69571	M F K FLVLTLG I I S C QAY …
ytcA	37.32636	M PTVLSRMAMQLK K TAWIIPVFM V S G C SLS …
yddW	36.63137	M DICSRNKKLTIR R PAILVALALL L C S C KST …

3.2 Comparison of various sources of lipoprotein predictions

The results of a variety of lipoprotein prediction programs and existing lipoprotein database entries were combined and corroborated to obtain a final compilation predicting a total of 125 lipoproteins in *E. coli* K-12. These were derived from a variety of sources and assessed for reliability. The LipoP training sets included 63 verified lipoproteins, 328 SPI-cleaved proteins, and 388 cytoplasmic proteins [10] selected in a variety of Gram-negative bacteria, *i.e.*, not just *E. coli*. Fifteen of the 63 verified lipoproteins in the LipoP positive training set were *E. coli* sequences. LipoP could detect an additional seven *E. coli* lipoproteins when SWISS-PROT protein sequences were used as opposed to those in GenBank [21], given that start sites were corrected in the SWISS-PROT version of the sequences [10].

In total, 134 *E. coli* SWISS-PROT and TrEMBL entries (release 42) are cross-linked to the PS000013 PROSITE pattern or the InterPro IPR000437 (signature characterizing bacterial lipoproteins) and 86 sequences are stored in DOLOP as the predicted set corresponding to *E. coli* K-12. The complete list of predicted lipoproteins reaches 101 items in [10]. The 81 proven lipoproteins listed in EcoGene (Table 1) do not exactly match data found in other databases. In SWISS-PROT, 78 of these *E. coli* entries are cross-linked with InterPro IPR000437. In DOLOP, 51 sequences are common to the set of 81. Furthermore, the published list of predicted proteins in [10] coincides only for 76/81 sequences used for training. Table 1 contains the citations to the experimental verification publications.

Table 5 lists all lipoproteins that were predicted by the various sources, excluding the entries in Table 1, reaching a total of 88 predicted, but as yet unproven, lipoproteins. In fact, 11/88 are proven not to be lipoproteins. Such recognized false positives are referenced in Table 5. One putative lipoprotein gene, *yahH*, listed in DOLOP only, has been removed from EcoGene as it is an unlikely translation of a REP element. Thirty-two out of 88 are evaluated as probable false positives based on homology analysis and other considerations listed in the 'Comments and False Positive References' column in Table 5. These are not infallible exclusionary considerations and some of these may in fact turn out to be lipoproteins, although this seems quite unlikely. This leaves 44/88 predicted as lipoproteins as shown in Table 3. Sixteen out of 44 possible lipoproteins did not have enough homologues to support the lipoprotein predictions. A conservative estimate for the predicted set, *i.e.*, probable lipoproteins, amounts to 28. Added to the verified lipoproteins (Table 1) the final *bona fide* set of lipoproteins contains a total of 109 proteins.

Table 5. Predicted *E. coli* K-12 lipoproteins

EG Acc	Gene	SP Acc	Len	Description	PATOSEQ/ LipoP	SP Annotation	DOLOP	Comments and False Positive References
LipoP and PATOSEQ hits								
EG12218	dcrB	P37620	185	Resistant to lytic phage C1	PATO/ LipoP-Yes	IPR000437	No	Probable lipoprotein, Cys conserved
EG11952	mdtP	P32714	488	Putative outer membrane factor for MdtNOP efflux pump	PATO/ LipoP-Yes	IPR000437, PS00013	Yes	Probable lipoprotein, Cys conserved
EG14380	rzoD	P58041	60	Homolog to lambda Rz1 lipoprotein in prophage DLP1 2	PATO/ LipoP-Yes	IPR000437, PS00013	No	Probable lipoprotein, Cys conserved, homologous to Rz1
EG14381	rzoR	P58042	61	Homolog to lambda Rz1 lipoprotein in prophage Rac	PATO/ LipoP-Yes	IPR000437, PS00013	No	Probable lipoprotein, Cys conserved, homologous to Rz1
EG10952	smpA	P23089	113	Putative lipoprotein, OmlA homolog, function unknown	PATO/ LipoP-Yes	IPR000437, PS00013	Yes	Probable lipoprotein, Cys conserved
EG12138	yaeF	P37056	274	Putative lipoprotein, function unknown	PATO/ LipoP-Yes	IPR000437, PS00013	No	Probable lipoprotein, Cys conserved
EG13608	yaiW	P77562	364	Putative lipoprotein, function unknown	PATO/ LipoP-Yes	IPR000437	Yes	Probable lipoprotein, Cys conserved
EG12182	yajG	P36671	192	Putative lipoprotein, function unknown	PATO/ LipoP-Yes	IPR000437, PS00013	No	Probable lipoprotein, Cys conserved

Table 5. Continued

EG Acc	Gene	SP Acc	Len	Description	PATOSEQ/ LipoP	SP Annotation	DOLOP	Comments and False Positive References
EG12874	yajI	P46122	179	Putative lipoprotein, function unknown	PATO/ LipoP-Yes	IPR000437, PS00013	No	Probable lipoprotein, Cys conserved
EG13430	ycfL	P75946	125	Putative lipoprotein, function unknown	PATO/ LipoP-Yes	IPR000437	Yes	Probable lipoprotein, Cys conserved
EG13182	ydbJ	P52646	88	Putative lipoprotein, function unknown	PATO/ LipoP-Yes	IPR000437	No	Probable lipoprotein, Cys conserved
EG14095	yfbK	P76481	575	Putative lipoprotein, function unknown	PATO/ LipoP-Yes	IPR000437	Yes	Probable lipoprotein, Cys conserved
EG13394	yfhM	P76578	1653	Putative lipoprotein, function unknown	PATO/ LipoP-Yes	IPR000437, PS00013	Yes	Probable lipoprotein, Cys conserved
EG12857	yfiM	P46126	107	Putative lipoprotein, function unknown	PATO/ LipoP-Yes	non-EcoGene start	No	Probable lipoprotein, Cys conserved
EG11291	yggG	P25894	252	Putative metalloprotease lipoprotein, function unknown	PATO/ LipoP-Yes	IPR000437, PS00013	No	Probable lipoprotein, Cys conserved
EG12994	yghJ	Q46837	1520	Putative lipoprotein, AcfD homolog, function unknown	PATO/ LipoP-Yes	IPR000437, PS00013	Yes	Probable lipoprotein, Cys conserved
EG12273	yiaF	P37667	236	Putative lipoprotein, function unknown	PATO/ LipoP-Yes	non-EcoGene start	No	Probable lipoprotein, Cys conserved
EG11719	yidX	P31461	218	Putative lipoprotein, function unknown	PATO/ LipoP-Yes	IPR000437	No	Probable lipoprotein, Cys conserved
EG12353	yifL	P39166	67	Putative lipoprotein, function unknown	PATO/ LipoP-Yes	IPR000437, PS00013	No	Probable lipoprotein, Cys conserved
EG12489	yjfO	P39297	109	Putative lipoprotein, function unknown	PATO/ LipoP-Yes	IPR000437, PS00013	No	Probable lipoprotein, Cys conserved
EG13719	ymbA	P75866	187	Putative lipoprotein, function unknown	PATO/ LipoP-Yes	non-EcoGene start	No	Probable lipoprotein, Cys conserved
EG12778	yraM	P45464	678	Putative lipoprotein, LppC homolog, function unknown	PATO/ LipoP-Yes	No	No	Probable lipoprotein, Cys conserved
EG13337	yafY	P77365	147	Function unknown	PATO/ LipoP-Yes	non-EcoGene start	No	Possible lipoprotein, yfjS is the only homolog
EG14001	ydjY	P76220	225	Function unknown	PATO/ LipoP-Yes	No	No	Possible lipoprotein, no homologs
EG14061	yegR	P76406	105	Function unknown	PATO/ LipoP-Yes	non-EcoGene start	No	Possible lipoprotein, no homologs
EG13205	yfjS	O52982	147	Function unknown	PATO/ LipoP-Yes	IPR000437, PS00013	No	Possible lipoprotein, yafY is the only homolog
EG14376	ypdI	O32528	91	Function unknown	PATO/ LipoP-Yes	IPR000437, PS00013	No	Possible lipoprotein, no homologs
EG10322	fliL	P06973	154	Affects rotational direction of flagella during chemotaxis	PATO/ LipoP-Yes	No	No	Probable false positive, Cys not conserved
PATOSEQ, not LipoP hits								
EG10178	cyoA	P18400	315	Cytochrome *c* oxidase subunit II, membrane-bound	LipoP-No	IPR000437, PS00013	No	Verified lipoprotein [32], 2 CM TMS predicted
EG14401	ytcA	None	91	Putative lipoprotein, function unknown	LipoP-No	Not in SWISS-PROT	No	Probable lipoprotein, lipobox conserved in Yersina
EG14316	yecT	P76296	162	Function unknown	LipoP-No	No	Yes	Possible lipoprotein, EC and distant homologs only

Table 5. Continued

EG Acc	Gene	SP Acc	Len	Description	PATOSEQ/ LipoP	SP Annotation	DOLOP	Comments and False Positive References
EG14093	arnT	P76473	550	4-amino-4-deoxy-L-arabinose: Lipid A transferase	LipoP-No	IPR000437	No	Probable false positive, 12 CM TMs predicted, Cys not conserved
EG10289	fecD	P15029	318	Ferric citrate transport membrane permease	LipoP-No	IPR000437	Yes	Probable false positive, 9 CM TMs predicted, Cys not conserved
EG10298	fepG	P23877	330	Ferrienterobactin permease, membrane-bound	LipoP-No	IPR000437	Yes	Probable false positive, 9 CM TMs predicted, Cys conserved in enterics
EG14077	rtn	P76446	518	Overexpression confers resistance to lambda and N4	LipoP-No	IPR000437	No	Probable false positive, 2 CM TMs predicted, Cys not conserved
EG11807	yebF	P33219	122	Function unknown	LipoP-No	IPR000437, PS00013	Yes	Probable false positive, Cys not conserved, probable SPI substrate
EG13473	yliB	P75797	512	Putative periplasmic binding protein, function unknown	LipoP-No	IPR000437, PS00013	No	Probable false positive, Cys not conserved, possible SPI substrate
EG14228	yqiG	P76655	822	Function unknown, fimbrial usher homolog	LipoP-No	non-EcoGene start	No	Probable false positive, Cys not conserved, IS21 insertion
EG10040	ampC	P00811	377	Intrinsic weak bet a-lactamase activity	LipoP-No	IPR000437	No	False positive, verified SPI substrate [55]
LipoP, not PATOSEQ hits								
EG12020	mdtQ	P33369	478	Putative OM lipoprotein of tripartite efflux pump	PATO-No	IPR000437, PS00013	No	Probable lipoprotein signal, Cys conserved, paralogs MdtP and CusC
EG12139	yfhG	P37328	237	Putative lipoprotein, function unknown	PATO-No	IPR000437	No	Probable lipoprotein signal, Cys is conserved
EG11712	yidQ	P31454	110	Putative lipoprotein, function unknown	PATO-No	IPR000437	No	Probable lipoprotein, Cys is conserved and yceK paralog is lipoprotein
EG11917	yjaH	P32681	231	Function unknown	PATO-No	No	No	Possible lipoprotein, Cys conserved, 30 aa long signal predicted
EG11926	yjbH	P32689	698	Function unknown, ymcA paralog	PATO-No	IPR000437, PS00013	No	Possible lipoprotein, Cys somewhat conserved
EG14400	ysaB	None	99	Function unknown	PATO-No	Not in SWISS-PROT	No	Possible lipoprotein, Cys conserved, no charged residue
EG11164	ygiB	P24195	223	Function unknown	PATO-No	non-EcoGene start	No	Possible lipoprotein signal, Cys is conserved, long signal (35aa)
EG12258	bcsZ	P37651	368	Endo-1,4-D-glucanase, periplasmic cellulase	PATO-No	No	No	Probable false positive, Cys somewhat conserved, soluble periplasmic

Table 5. Continued

EG Acc	Gene	SP Acc	Len	Description	PATOSEQ/ LipoP	SP Annotation	DOLOP	Comments and False Positive References
EG14240	kefA	P77338	1120	Mechanosensitive channel protein MscK(KefA)	PATO-No	No	No	Probable false positive, 12 CM TMs, Cys not conserved
EG11950	nrfG	P32712	198	Required for Nrf pathway, function unknown,	PATO-No	No	No	Probable false positive, Cys not conserved
EG11333	visC	P25535	400	Putative FAD-dependent oxidoreductase, function unknown	PATO-No	No	No	Probable false positive, Cys not conserved, probably cytoplasmic
EG11769	ybbC	P33668	122	Function unknown	PATO-No	IPR000437, PS00013	No	Probable false positive, Cys not conserved
EG13650	ybeT	P77296	184	Function unknown	PATO-No	No	No	Probable false positive, Cys not conserved
EG11780	ydeK	P32051	1325	Putative OM autotransporter adhesin, function unknown	PATO-No	IPR000437, PS00013	No	Probable false positive, Cys not conserved in other autotransporters
EG11840	yihN	P32135	421	Putative MFS family permease, function unknown	PATO-No	No	No	Probable false positive, 10 CM TMs, Cys not conserved
EG10530	lepB	P00803	324	Signal peptidase I (for nonlipoproteins)	PATO-No	No	No	False positive, shown to have no signal peptide [58]
EG13271	panE	P77728	303	Ketopantoate reductase, NADPH-dependent	PATO-No	No	No	False positive, unprocessed, verified by mass spectrometry [62]
EG10971	srlD	P05707	259	Sorbitol-6-phosphate dehydrogenase	PATO-No	No	No	False positive, unprocessed, verified amino terminus [43]
Swiss-Prot/InterPro only								
EG13149	yafL	Q47151	249	Putative lipoprotein, function unknown	PATO/ LipoP-No	IPR000437	No	Probable lipoprotein, NlpC paralog
EG11488	ydhA	P28224	109	Putative lipoprotein, function unknown	PATO/ LipoP-No	IPR000437	No	Probable lipoprotein, lipobox conserved, EcoGene had unlikely start codon
EG14383	mgrB	P76267	47	Mg(2+)-starvation-stimulated gene, function unknown	PATO/ LipoP-No	IPR000437	No	Possible lipoprotein, Cys conserved in Salmonella
EG13477	yliF	P75801	442	Function unknown	PATO/ LipoP-No	IPR000437	No	Possible lipoprotein, no N-domain homologs
EG13729	ymcA	P75882	698	Function unknown, yjbH paralog	PATO/ LipoP-No	IPR000437, PS00013	No	Possible lipoprotein, Cys somewhat conserved
EG14007	ynjE	P78067	435	Rhodanese-like protein, function unknown	PATO/ LipoP-No	IPR000437	No	Possible lipoprotein, Cys conserved
EG13178	rseC	P46187	159	Required for the reduction of SoxR	PATO/ LipoP-No	IPR000437	Yes	Probable false positive, lipobox not conserved
EG13643	ybdJ	P77506	82	Function unknown	PATO/ LipoP-No	IPR000437	No	Probable false positive, Cys not conserved, 2 CM TMs predicted

Table 5. Continued

EG Acc	Gene	SP Acc	Len	Description	PATOSEQ/ LipoP	SP Annotation	DOLOP	Comments and False Positive References
EG12395	ybgE	P37343	97	Fourth gene in cydAB operon, function unknown	PATO/ LipoP-No	IPR000437	No	Probable false positive, Cys not conserved
EG13312	ybgP	P75749	242	Putative periplasmic pilus chaperone, function unknown	PATO/ LipoP-No	IPR000437	No	Probable false positive, Cys not conserved, likely SPI substrate
EG13710	ycbR	P75856	233	Putative periplasmic pilus chaperone, function unknown	PATO/ LipoP-No	IPR000437	No	Probable false positive, Cys not conserved, likely SPI substrate
EG11735	ycdB	P31545	423	Function unknown, peroxidase homolog	PATO/ LipoP-No	IPR000437, PS00013	No	Probable false positive, Cys not conserved, predicted Tat substrate
EG13970	ydiK	P77175	370	Putative membrane permease, function unknown	PATO/ LipoP-No	IPR000437	Yes	Probable false positive, Cys not conserved, 9 CM TMs predicted
EG14164	yfeW	P77619	434	Putative periplasmic esterase, function unknown	PATO/ LipoP-No	IPR000437	No	Probable false positive, Cys not conserved, likely SPI substrate
EG10018	yhdA	P13518	646	Function unknown	PATO/ LipoP-No	IPR000437, PS00013	No	Probable false positive, Cys not conserved
EG11267	yiaB	P11286	113	Inner membrane protein, function unknown	PATO/ LipoP-No	IPR000437	No	Probable false positive, Cys not conserved, 4 CM TMs predicted
EG12281	yiaM	P37674	157	Putative membrane permease, function unknown	PATO/ LipoP-No	IPR000437	No	Probable false positive, Cys not conserved, 4 CM TMs predicted
EG14229	yqiH	P77616	249	Putative periplasmic pilus chaperone, function unknown	PATO/ LipoP-No	IPR000437	No	Probable false positive, Cys not conserved, likely SPI substrate
EG10315	fimH	P08191	300	Minor type 1 fimbrial adhesion subunit	PATO/ LipoP-No	IPR000437	No	False positive, verified SPI substrate [56]
EG10374	ggt	P18956	580	gamma-Glutamyltranspeptidase	PATO/ LipoP-No	IPR000437	No	False positive, Verified SPI substrate [52]
DOLOP only								
EG12816	nanE	P45426	229	Putative ManNAc-6-Pto GlcNAc-6-P epimerase	PATO/ LipoP-No	No	Yes	Possible lipoprotein, Cys conserved
EG14386	yaaY	P75620	72	Function unknown	PATO/ LipoP-No	No	Yes	Possible lipoprotein, EC-ST only
EG11052	uhpB	P09835	500	Membrane protein controlling UhpA activity, sensor kinase	PATO/ LipoP-No	No	Yes	Probable false positive, Cys not conserved, 10 CM TMs predicted
EG12097	yfiH	P33644	243	Function unknown	PATO/ LipoP-No	No	Yes	Probable false positive, Cys not conserved
EG14163	yfeV	P77272	474	Putative PTS system IIBC component, function unknown	PATO/ LipoP-No	No	Yes	Probable false positive, Cys not conserved, 9 CM TMs predicted
EG13003	yghS	Q46843	237	Function unknown	PATO/ LipoP-No	No	Yes	Probable false positive, Cys not conserved
EG13839	ynfA	P76169	108	Inner membrane protein, function unknown	PATO/ LipoP-No	No	Yes	False positive (K.E.R., unpublished)

Table 5. Continued

EG Acc	Gene	SP Acc	Len	Description	PATOSEQ/ LipoP	SP Annotation	DOLOP	Comments and False Positive References
EG10120	bioD	P13000	225	Dethiobiotin synthase	PATO/ LipoP-No	No	Yes	False positive, verified amino terminus, Met is cleaved [59]
EG10202	dacB	P24228	477	D-alanine D-alanine carboxypeptidase PBP4	PATO/ LipoP-No	No	Yes	False positive, verified SPI substrate [60]
EG10306	fhuE	P16869	729	Outer membrane receptor for ferric-rhodotorulic acid	PATO/ LipoP-No	No	Yes	False positive, verified SPI substrate [61]
EG11481	tatD	P27859	260	Mg-dependent cytoplasmic DNase	PATO/ LipoP-No	No	Yes	False positive, unprocessed, verified amino terminus [57]
EG13592	yahH	P75690	106	YahH is no longer in EcoGene	PATO/ LipoP-No	No	Yes	Defunct gene, unlikely translation of REP sequences

4 Discussion

Automatic classification of protein sequences depends, in most cases, on the presence of patterns and motifs. Motifs are generally determined as regions of conserved positions in the optimized alignment of amino acid sequences. In other words, regularities identified in a conserved region are expressed as positional constraints. Such denoted consensus sequences often correspond to binding sites for substrates or regions involved in modification, transport, degradation, *etc*. In protein regions identified as cleavage sites in protein processing pathways, even though amino acid regularity is visually obvious, the variability of sequence length affects the quality of alignment through the introduction of a substantial number of gaps. Furthermore, many of the features of protein binding sites are given in terms of characteristics, such as net charge, hydrophobicity, size, *etc*., which are not necessarily well represented by the presence or absence of a specific amino acid residue.

Such problems are exemplified in the case of the cleavable *N*-terminal regions of bacterial proteins. Searches for signal peptide patterns, irrespective of their type, were formalized along three main guidelines. Regular expressions were used to accommodate length variability but their implementation usually generates a binary answer (presence/absence of a motif) [19]. Alternatively, strategies using neural nets were defined to provide scoring functions but users are deprived from an explicit and rational biological explanation for an output [10, 22]. Rule-based systems [23] ideally circumvent the cited shortcomings associated with the use of regular expressions and neural nets but can only reproduce the limitations of human understanding.

PATOSEQ introduced in [9] was set as an attempt to identify further explicit rules and constraints that might not only be positional and would reflect a biological phenomenon. We first suggested to change the alphabet used for describing motifs in order to include partial information on positional constraints in the descriptors. Secondly, we set interdependent matching and scoring procedures that would guarantee stable and optimized scores. Given a motif description, scoring was set as the maximized probability for a sequence to match this description. But, as mentioned early in Section 2.1, in all cases, the most critical step remains the initial selection of a reference or a training set.

In the framework of bacterial lipoprotein study, attention has first been focussed on the consensus defining the so-called lipobox as initially identified in [5, 6]. The presence of this consensus has set the basis of all patterns used for lipoprotein recognition. At the time, much fewer sequences were available than nowadays. Incoming genome data spurred further direct investigations of sequence patterns with *ad hoc* methods in Gram-positive bacteria [24, 25] as well as all bacteria indiscriminately [12]. The latter updated resource provides a looser definition of the PROSITE pattern for searching potential lipoprotein signal sequence that allows a high number of false positives. The regular expression defined as the PROSITE pattern and complemented with *ad hoc* rules is more stringent.

We considered the most documented bacterial proteome, *i.e.* that of *E. coli* K-12. An initial set of 81 lipoproteins was carefully checked and crosschecked for maximum guarantee to comply with the standards of annotation in EcoGene [13]. It is justified in Section 3. This set

was used for training PATOSEQ to search lipoproteins. In parallel, 134 SWISS-PROT and TrEMBL entries cross-linked to the PS000013 PROSITE pattern or the InterPro IPR000437 were retrieved as well as the 86 predicted lipoproteins of *E. coli* K-12 in DOLOP. The variability of sources (EcoGene, SWISS-PROT, DOLOP, InterPro and PROSITE) and the differences of the prediction schemes provided by LipoP and PATOSEQ motivated the validation of each sequence.

Unsurprisingly, the quality of annotation is uneven and proportional to the level of human input. In particular, improper protein starts are a significant cause for inconsistencies that bear on the accuracy of database information as well as the performance of predictive methods. Indeed, feeding LipoP with EcoGene protein sequences that included recently revised predicted translation start sites, relative to GenBank or SWISS-PROT, enhanced the efficiency of the program: six additional lipoproteins were predicted (noted in the LipoP predictions of Table 5 as "non-EcoGene starts"). The EcoGene start site prediction revisions are presented in the EcoGene records for these genes in the "Gene Quality" field (http://bmb.med.miami.edu/EcoGene/EcoWeb).

Each item of the predicted lipoproteins in *E. coli* K-12 listed in this paper was analyzed in detail. Explanations correspond to the best of our knowledge *via* the use of heuristic rules (conservation in orthologues, prediction of transmembrane regions and other topological criteria). Using the corrected start sites from EcoProt, LipoP and PATOSEQ yield 109 matching predictions. Quite logically, PATOSEQ recognized the 81 proven lipoproteins. It also predicted another 37, 28 of which we consider to be correct predictions. The remaining nine predictions which we consider to be false are either proven or supposed Type I secreted proteins. It should be noted that these are not selected by the LipoP predictor since a filter for Type I signals is applied prior to processing sequences for Type II.

The explicit performance of PATOSEQ helped identify some specific features of the lipoprotein signal in *E. coli* K-12, proper. In particular the nonappearance of particular amino acids in the lipobox influences the prediction and may provide further constraints justifying protein secretion. Conversely, PATOSEQ is inflexibly sensitive to the absence of positively charged residues between the initial methionine and the helical part of the signal. The investigation of apparently minute discrepancies in signal peptide sequences can lead to a more precise recognition. In fact, unpublished tests with slight variations on the motif in Gram-positive bacteria led to fine-tune descriptions depending on the organism. Such small differences matched published observations in [25]. Given the physiological differences between organisms, variations in the properties of secretion are to be expected. We looked into characteristics that would distinguish pathogenic from nonpathogenic strains that could not be tested so far.

5 Concluding remarks

The careful sorting of *E. coli* K-12 lipoproteins led to the selection of a restricted set of candidates that could be tested. It also provided a benchmark set for evaluating predictive methods and their sensitivity. As a common trend in bioinformatics applications, the combined use of several methods is equivalent to merging several viewpoints; most of the time, it reinforces the reliability of prediction and shows that various methods complement each other.

We would like to thank Russell Bishop for his valuable help in compiling the data sets used in this work. We would also like to thank the editors for their infinite patience and the reviewers for their constructive comments. Part of this work was supported by the NIH grant No. GM58560 to KER.

6 References

[1] Gattiker, A., Michoud, K., Rivoire, C., Auchincloss, A. H. *et al.*, *Comput. Biol. Chem.* 2003, *27*, 49–58.
[2] Haft, D. H., Selengut, J. D., White, O., *Nucleic Acids Res.* 2003, *31*, 371–373.
[3] Perriere, G., Duret, L., Gouy, M., *Genome Res.* 2000, *10*, 379–385.
[4] Tobes, R., Ramos, J. L., *Nucleic Acids Res.* 2002, *30*, 318–321.
[5] Klein, P., Somorjai, R. L., Lau, P. C., *Protein Eng.* 1988, *2*, 15–20.
[6] von Heijne, G., *Protein Eng.* 1989, *2*, 531–534.
[7] Gattiker, A., Gasteiger, E., Bairoch, A., *Appl. Bioinformatics* 2002, *1*, 107–108.
[8] Zdobnov, E. M., Apweiler, R., *Bioinformatics* 2001, *17*, 847–848.
[9] Gonnet, P., Lisacek, F., *Bioinformatics* 2002, *18*, 1091–1101.
[10] Juncker, A. S., Willenbrock, H., von Heijne, G., Brunak, S. *et al.*, *Protein Sci.* 2003, *12*, 1652–1662.
[11] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C. *et al.*, *Nucleic Acids Res.* 2003, *31*, 365–370.
[12] Madan Babu, M., Sankaran, K., *Bioinformatics* 2002, *18*, 641–643.
[13] Rudd, K. E., *Nucleic Acids Res.* 2000, *28*, 60–64.
[14] Hayashi, S., Hara, H., Suzuki, H., Hirota, Y., *J. Bacteriol.* 1988, *170*, 5392–5395.
[15] Blattner, F. R., Plunkett, G. 3rd, Bloch, C. A., Perna, N. T. *et al.*, *Science* 1997, *277*, 1453–1474.
[16] Mathews, B. W., *Biochim. Biophys. Acta* 1975, *405*, 442–451.
[17] Fisher, R. A., *Ann. Eugenics* 1975, *7*, 179–188.

[18] Gonnet, G. H., LinearClassify 2001, http://cbrg.ethz.ch/Darwin.

[19] Bairoch, A., *Nucleic Acids Res.* 1991, *19*, 2241–2245.

[20] Gibbs, M. L., Jacobs, M., Wilkie, A. O., Taylor, D., *J. Pediatr. Ophthalmol. Strabismus* 1995, *32*, 142.

[21] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. *et al.*, *Nucleic Acids Res.* 2003, *31*, 23–27.

[22] Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G., *Protein Eng.* 1997, *10*, 1–6.

[23] Bannai, H., Tamada, Y., Maruyama, O., Nakai, K. *et al.*, *Bioinformatics* 2002, *18*, 298–305.

[24] Tjalsma, H., Bolhuis, A., Jongbloed, J. D., Bron, S. *et al.*, *Microbiol. Mol. Biol. Rev.* 2000, *64*, 515–547.

[25] Sutcliffe, I. C., Harrington, D. J., *Microbiology* 2002, *148*, 2065–2077.

[26] Zgurskaya, H. I., Nikaido, H., *J. Mol. Biol.* 1999, *285*, 409–420.

[27] Ichikawa, J. K., Li, C., Fu, J., Clarke, S., *J. Bacteriol.* 1994, *176*, 1630–1638.

[28] Beck, B. J., Downs, D. M., *J. Bacteriol.* 1998, *180*, 885–891.

[29] Bishop, R. E., Penfold, S. S., Frost, L. S., Holtje, J. V. *et al.*, *J. Biol. Chem.* 1995, *270*, 23097–23103.

[30] Barondess, J. J., Beckwith J., *J. Bacteriol.* 1995, *177*, 1247–1253.

[31] Loferer, H., Hammar, M., Normark, S., *Mol. Microbiol.* 1997, *26*, 11–23.

[32] Ma, J., Katsonouri, A., Gennis, R. B., *Biochemistry* 1997, *36*, 11298–11303.

[33] Bishop, R. E., Leskiw, B. K., Hodges, R. S., Kay, C. M. *et al.*, *J. Mol. Biol.* 1998, *280*, 583–596.

[34] Kraft, A. R., Templin, M. F., Holtje, J. V., *J. Bacteriol.* 1998, *180*, 3441–3477.

[35] Schoenhals, G. J., Macnab, R. M., *J. Bacteriol.* 1996, *178*, 4200–4207.

[36] Matsuyama, S., Yokota, N., Tokuda, H., *EMBO J.* 1997, *16*, 6947–6955.

[37] Braun, V., Bosch, V., *Eur. J. Biochem.* 1972, *28*, 51–69.

[38] Lommatzsch, J., Templin, M. F., Kraft, A. R., Vollmer, W. *et al.*, *J. Bacteriol.* 1997, *179*, 5465–5470.

[39] Ehlert, K., Holtje, J. V., Templin, M. F., *Mol. Microbiol.* 1995, *16*, 761–768.

[40] Yu, F., Inouye, S., Inouye, M., *J. Biol. Chem.* 1986, *261*, 2284–2288.

[41] Bouvier, J., Pugsley, A. P., Stragier, P., *J. Bacteriol.* 1991, *173*, 5523–5531.

[42] Hantke, K., Braun, V., *Eur. J. Biochem.* 1973, *34*, 284–296.

[43] Novotny, M. J., Reizer, J., Esch, F., Saier, M. H. Jr., *J. Bacteriol.* 1984, *159*, 986–990.

[44] Ohara, M., Wu, H. C., Sankaran, K., *J. Bacteriol.* 1999, 181, 4318–4325.

[45] Jung, J. U., Gutierrez, C., Villarejo, M. R., *J. Bacteriol.* 1989, *171*, 511–520.

[46] Mizuno, T., *J. Biochem. (Tokyo)* 1979, *86*, 991–1000.

[47] Chen, R., Henning, U., *Eur. J. Biochem.* 1987, *163*, 73–77.

[48] Takase, I., Ishino, F., Wachi, M., Kamata, H. *et al.*, *J. Bacteriol.* 1987, *169*, 5692–5699.

[49] Alexander, D. M., St John, A. C., *Mol. Microbiol.* 1994, *11*, 1059–1071.

[50] Ludwig, A., Tengel, C., Bauer, S., Bubert, A. *et al.*, *Mol. Gen. Genet.* 1995, *249*, 474–486.

[51] Drummelsmith, J., Whitfield, C., *EMBO J.* 2000, *19*, 57–66.

[52] Suzuki, H., Kumagai, H., Echigo, T., Tochikura, T., *J. Bacteriol.* 1989, *171*, 5169–5172.

[53] Seiffer, D., Klein, J. R., *FEMS Microbiol. Lett.* 1993, *107*, 175–178.

[54] Snyder, W. B., Davis, L. J., Danese, P. N., Cosma, C. L. *et al.*, *J. Bacteriol.* 1995, *177*, 4216–4223.

[55] Jaurin, B., Grundstrom, T., *Proc. Natl. Acad. Sci. USA* 1981, *78*, 4897–4901.

[56] Hanson, M. S., Hempel, J., Brinton, C. C. Jr., *J. Bacteriol.* 1988, *170*, 3350–3358.

[57] Wexler, M., Sargent, F., Jack, R. L., Stanley, N. R. *et al.*, *J. Biol. Chem.* 2000, *275*, 16717–16722.

[58] Wolfe, P. B., Wickner, W., Goodman, J. M., *J. Biol. Chem.* 1983, *258*, 12073–12080.

[59] Alexeev, D., Bury, S. M., Boys, C. W., Turner, M. A. *et al.*, *J. Mol. Biol.* 1994, *235*, 774–776.

[60] Mottl, H., Terpstra, P., Keck, W., *FEMS Microbiol. Lett.* 1991, *62*, 213–220.

[61] Sauer, M., Hantke, K., Braun, V., *J. Bacteriol* 1987, *169*, 2044–2049.

[62] Zheng, R., Blanchard, J. S., *Biochemistry* 2000, *39*, 3708–3717.

