

Research article

Open Access

## Genome-wide subcellular localization of putative outer membrane and extracellular proteins in *Leptospira interrogans* serovar Lai genome using bioinformatics approaches

Wasna Viratyosin\*<sup>1</sup>, Supawadee Ingsriswang<sup>1</sup>, Eakasit Pacharawongsakda<sup>1</sup> and Prasit Palittapongarnpim<sup>1,2</sup>

Address: <sup>1</sup>BIOETC Central Research Unit, National Center for Genetic Engineering and Biotechnology, Pathumthani, 12120, Thailand and <sup>2</sup>Department of Microbiology, Faculty of Science, Mahidol University, Bangkok, 10400, Thailand

Email: Wasna Viratyosin\* - wasna@biotec.or.th; Supawadee Ingsriswang - supawadee@biotec.or.th; Eakasit Pacharawongsakda - eakasit.pac@biotec.or.th; Prasit Palittapongarnpim - Prasit@biotec.or.th

\* Corresponding author

Published: 21 April 2008

Received: 4 December 2007

BMC Genomics 2008, **9**:181 doi:10.1186/1471-2164-9-181

Accepted: 21 April 2008

This article is available from: <http://www.biomedcentral.com/1471-2164/9/181>

© 2008 Viratyosin et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** In bacterial pathogens, both cell surface-exposed outer membrane proteins and proteins secreted into the extracellular environment play crucial roles in host-pathogen interaction and pathogenesis. Considerable efforts have been made to identify outer membrane (OM) and extracellular (EX) proteins produced by *Leptospira interrogans*, which may be used as novel targets for the development of infection markers and leptospirosis vaccines.

**Result:** In this study we used a novel computational framework based on combined prediction methods with deduction concept to identify putative OM and EX proteins encoded by the *Leptospira interrogans* genome. The framework consists of the following steps: (1) identifying proteins homologous to known proteins in subcellular localization databases derived from the "consensus vote" of computational predictions, (2) incorporating homology based search and structural information to enhance gene annotation and functional identification to infer the specific structural characters and localizations, and (3) developing a specific classifier for cytoplasmic proteins (CP) and cytoplasmic membrane proteins (CM) using Linear discriminant analysis (LDA). We have identified 114 putative EX and 63 putative OM proteins, of which 41% are conserved or hypothetical proteins containing sequence and/or protein folding structures similar to those of known EX and OM proteins.

**Conclusion:** Overall results derived from the combined computational analysis correlate with the available experimental evidence. This is the most extensive *in silico* protein subcellular localization identification to date for *Leptospira interrogans* serovar Lai genome that may be useful in protein annotation, discovery of novel genes and understanding the biology of *Leptospira*.

## Background

Leptospirosis is a globally widespread zoonosis caused by the animal spirochete pathogen *Leptospira interrogans* [1]. The clinical feature of its severe disease form, known as Weil's syndrome, or acute renal failure, is associated with multiple system complications, including renal failure, meningitis, and pulmonary haemorrhage. Although early treatment for leptospirosis is important for ensuring a favorable clinical outcome, this is often difficult to achieve, as symptoms during the early stages of infection resemble those of several other systematic diseases.

One potential method for controlling the spread of leptospirosis is through the development of vaccines. Candidates for vaccine production include outer membrane (OM) and extracellular (EX) proteins, several of which have been implicated in chemotaxis, adherence and other pathogenic steps. Attempts to identify such proteins have been performed previously by experimental [2-14] and computational methods [15-20]. Complete genome sequences of two serovars, Lai and Copenhageni of *L. interrogans* have been reported [15-17]. Hundreds of putative membrane proteins and lipoproteins were predicted, although in many cases, gene annotation may be incomplete or inaccurate to reliably identify putative vaccine candidates.

Previous studies have tried to identify potential vaccine candidates using experimental methods and *in silico* predictions. Proteomic analysis of purified outer membrane vesicles (OMVs) of *L. interrogans* serovar Copenhageni was performed by Nally *et al.* and revealed 33 intact OM proteins [13]. The study by Gamberini *et al.* [18] showed 16 predicted surface exposed lipoproteins of *L. interrogans* serovar Copenhageni via whole genome analysis, only four of which are conserved among 8 pathogenic serovars. Since leptospiral lipoproteins are usually (but not exclusively) surface exposed proteins, and many are vaccine candidates, Setubal *et al.* [19] focused on lipoprotein prediction using spirochaetal lipoprotein (SpLip) program and identified 146 predicted lipoproteins (but not their localizations) for *L. interrogans* serovar Lai. The search for new potential vaccine candidates was continued by Yang *et al.* [20], who used a filtering approach combining *in silico* analysis, comparative genome hybridization, and microarray methods to identify 226 leptospiral surface exposed proteins. All of the previous studies summarized above focus on identification of vaccine candidates.

However, both computational and experimental have their own drawbacks [21,22]. Computational methods, for instance, depend on the presence of type I signal peptides [23,24], transmembrane helices [24-26], or other particular features specifically found in previously identified membrane proteins, which may not be highly specific

or sensitive. Experimental methods, on the other hand, yield results that may be complicated by cross-compartment contamination occurring during the preparation of samples, which can also result in the inclusion of false positive results in data sets [21,22]. Hence, results obtained from both methods can occasionally lead to conflicting conclusions. We believe that such a focused approach without attempt to accurately identify periplasmic proteins (PP) and cytoplasmic membrane (CM) proteins can lead to erroneous identification of PP and CM as OM or EX by both *in silico* and experimental approaches. A holistic prediction of all membrane protein localizations will lead to better accuracy in genome annotation of membrane proteins, including vaccine candidates.

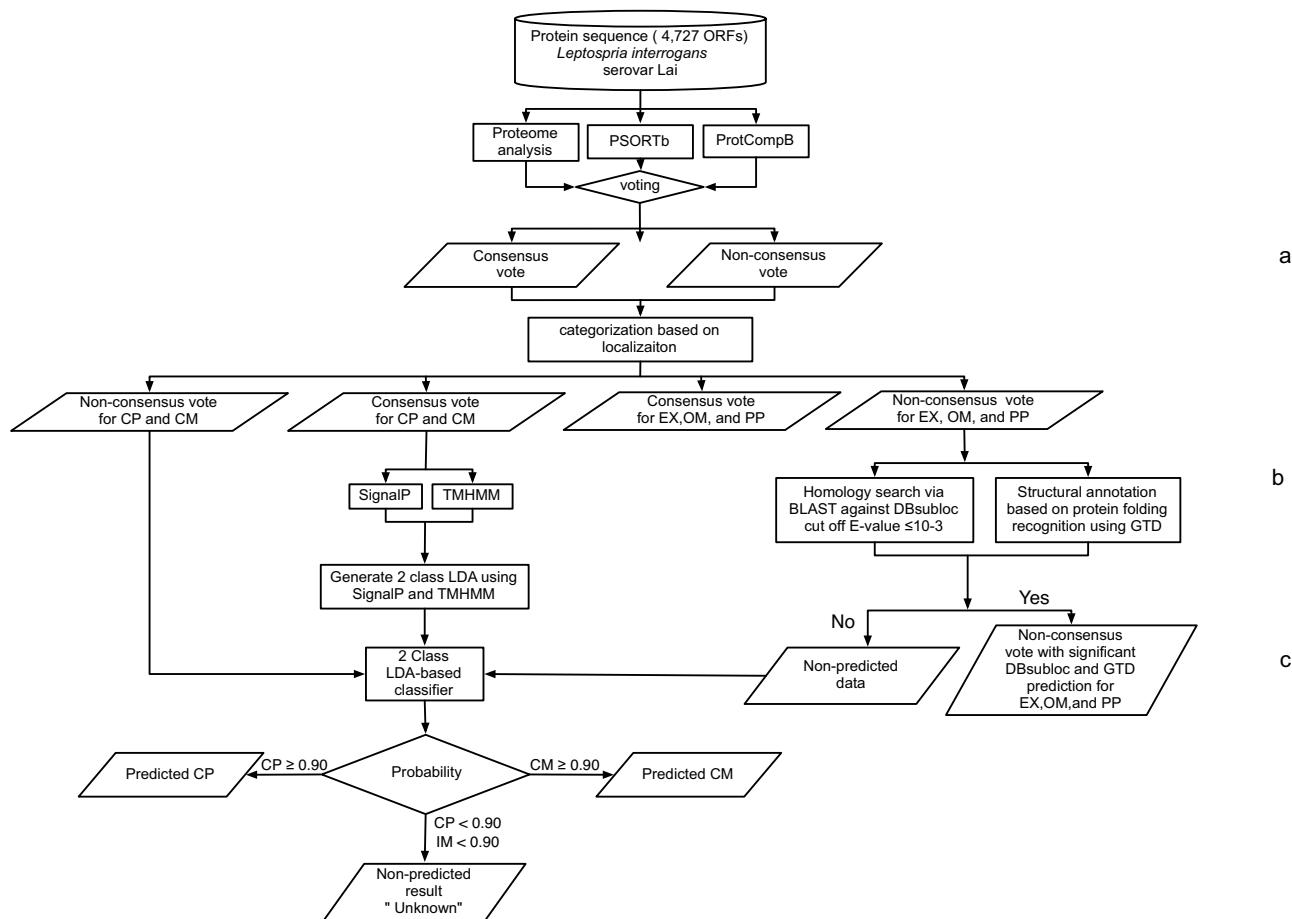
In this study we utilized a combination of three computational prediction tools PSORTb [27,28], Proteome Analyst (PA) [29], and ProtCompB [30] to perform whole genome analysis of protein subcellular localization, and to identify novel putative *L. interrogans* serovar Lai OM and EX vaccine candidates. We combined the results derived from these three prediction algorithms into a consensus vote, resulting in a more accurate protein subcellular localization prediction. Furthermore, we incorporated homology searching against the DBSubloc database [31] and structural information from the GTD prediction [32] to enhance genome annotation, and to infer OM, EX and PP localized proteins. We also developed a specific classifier based on Linear Discriminant Analysis (LDA) for identification of leptospiral cytoplasmic proteins (CP) and cytoplasmic membrane proteins (CM), using a training set obtained from the consensus vote. We were able to assign subcellular localizations to several previously uncharacterized hypothetical proteins, thus improving *L. interrogans* genome annotation.

## Results

We performed the subcellular localization prediction of *L. interrogans* serovar Lai using the pipeline described in the Material and methods section (shown in Figure 1), following the steps of training set verification, consensus vote, homology and structural prediction, and finally LDA-based classification.

### **Training set verification: Localization predictions of a set of experimentally verified proteins with known localization**

To evaluate the robustness and versatility of our protein localization procedure, we used a set of well-characterized Gram-negative bacterial proteins with experimentally verified localizations taken from the work by Gardy and Brinkman [22] as a test set. The data set comprising 299 proteins was first analyzed by using PSORTb, PA, and ProtCompB. We found that, individually, PSORTb, PA, and ProtCompB assigned 73%, 71% and 79% of the verified

**Figure 1**

**Flow chart of the method used for subcellular localizations of *Leptospira interrogans* serovar Lai genome.** Protein sequences of *Leptospira interrogans* serovar Lai genome (4,727 ORFs) were analyzed for subcellular localization using PSORTb, ProtCompB, and Proteome analyst (PA) prediction. (a) The consensus vote was obtained from the majority vote type procedure to obtain the result with high prediction accuracy. If all 3 methods agree for localization it was assigned as a consensus vote. The remaining (1 or 2 out of 3 predicted result) was assigned as non-consensus vote. The consensus vote of CP and CM was used as a training set for the development of an LDA-based classifier for CP and CM in the next step. (b) The non-consensus vote results of OM, PP, and EX were further analyzed for sequence and structure homology by DBSubloc and GTD prediction. The non-consensus vote of EX, OM, and PP with significant homology or/and structure information were identified by DBSubloc and GTD prediction. (c) Non-consensus votes of CP, CM and the non predicted data from DBSubloc and GTD predictions were further analyzed for subcellular localization using LDA-based classifier for CP and CM. Significantly predicted results were proteins classified with more than 0.90 probability for CP and CM proteins. The remaining queries that could not be identified in this step were classified as "unknown" results.

protein localizations respectively (recall rate in Table 1). The overall precision rates were 97%, 95 and 83%, respectively. As expected, the overall recall rate was highest for ProtCompB, while its precision rate was also the lowest. The recall rate based on "consensus vote" (see materials and methods) results derived from all three methods was 48% without any false positives. Relaxing the criteria by considering predicted results of any two methods or the "majority vote" resulted in an overall recall rate of 77% with a single false positive.

Since the number of outputs for EX and OM proteins agreed by all three predictions was low (low recall rate), we used structure-based homology information from GTD and/or homology search results from DBSubloc prediction as the additional information for inferring protein localization. Using this information, we assessed the likelihood of the "non-consensus vote" outputs (see material and methods) for being EX or OM proteins. When the information from DBSubloc and GTD predictions were also used, the overall recall rates for the EX, OM and PP

**Table I: Localization predictions of a set of 299 experimentally verified proteins with known localization**

Actual localization	Total	TP	FP	FN	TN	Precision	Recall
<b>PSORTb</b>							
CP	145	110	1	35	110	99.10%	75.86%
CM	69	55	2	14	166	96.49%	79.71%
PP	29	18	0	11	207	100.00%	62.07%
OM	38	30	0	8	195	100.00%	78.95%
EX	18	6	3	12	216	66.67%	33.33%
Total	299	219	6	80	894	97.33%	73.24%
<b>Proteome Analyst</b>							
CP	145	94	0	51	119	100.00%	64.83%
CM	69	59	2	10	162	96.72%	85.51%
PP	29	19	3	10	201	86.36%	65.52%
OM	38	31	0	7	192	100.00%	81.58%
EX	18	9	6	9	207	60.00%	50.00%
Total	299	212	11	87	881	95.07%	70.90%
<b>ProtCompB</b>							
CP	145	127	11	18	144	92.03%	87.59%
CM	69	55	7	14	227	88.71%	79.71%
PP	29	19	9	10	261	67.86%	65.52%
OM	38	23	18	15	243	56.10%	60.53%
EX	18	11	4	7	277	73.33%	61.11%
Total	299	235	49	64	1152	82.75%	78.60%
<b>Consensus vote</b>							
CP	145	67	0	78	154	100.00%	46.21%
CM	69	43	0	26	230	100.00%	62.32%
PP	29	11	0	18	270	100.00%	37.93%
OM	38	19	0	19	261	100.00%	50.00%
EX	18	4	0	13	216	100.00%	23.53%
Total	299	144	0	154	1131	100.00%	48.32%
<b>Majority vote (2 out of 3 predictions)</b>							
CP	145	121	0	24	154	100.00%	83.45%
CM	69	59	0	10	230	100.00%	85.51%
PP	29	17	0	12	270	100.00%	58.62%
OM	38	29	0	10	213	100.00%	74.36%
EX	18	6	1	12	215	85.71%	33.33%
Total	299	232	1	68	1082	99.57%	77.33%
<b>Combination method</b>							
CP	N/A	N/A	N/A	N/A	N/A	N/A	N/A
CM	N/A	N/A	N/A	N/A	N/A	N/A	N/A
PP	29	25	0	4	56	100.00%	86.20%
OM	38	34	1	4	46	97.14%	89.47%
EX	18	12	2	6	65	85.71%	66.67%
Total	85	71	3	14	167	95.95%	87.53%

299 proteins obtained from the test set used in comparison study by Gardy and Brinkman [22]. Majority vote is the result from 2 out of 3 predictions. Combination method: the result from non-consensus vote with significant DBSubloc [31] and/or GTD prediction [32]. Precision is calculated as TP/(TP+FP), Recall is calculated as TP/(TP+FN) TP = true positive, TN = true negative, FP = false positive, FN = false negative, N/A = Not applicable

increased to 67%, 89% and 86% respectively as shown in Table 1. The method resulted in 96% precision. This performance was much better than any of the three individual methods, or any of the above combinations. Therefore, we have shown that the combination of prediction tools, DBSubloc homology search and GTD structural-based prediction markedly improved the accuracy and recall for EX, OM and PP protein localization prediction. Therefore, our prediction pipeline is applicable for

subcellular localization prediction of hypothetical, or unknown proteins.

#### **Subcellular localization predictions of *L. interrogans*: Step 1 Consensus votes**

After demonstration of the accuracy of our pipeline prediction with the training set, the whole predicted proteome of *L. interrogans* serovar Lai was analyzed using three computational predictions for protein subcellular

**Table 2: Predicted protein subcellular localizations of *L. interrogans* by PSORTb, PA, ProtCompB and consensus vote predictions.**

Localization	Subcellular localization prediction			
	PSORTb	PA	ProtCompB	Consensus vote
Cytoplasm (CP)	1125	921	2013	418
Cytoplasmic membrane (CM)	606	715	1726*	332
Outer membrane (OM)	112	28	-	15
Periplasmic (PP)	30	86	478	17
Extracellular (EX)	29	326	510	15
Unknown	2825	2652	-	3930

\* Note that ProtCompB prediction in this version, CM and OM were predicted as membrane proteins.

localization: PSORTb, ProtCompB, and Proteome analyst (PA). The results obtained from each prediction program are shown in Table 2. ProtCompB assigned subcellular localizations to all protein queries whereas approximately 50% of protein queries were assigned as unknown localization by PSORTb and PA.

After inspection of the prediction results derived from the three prediction algorithms, it was found that 797 out of 4,727 ORFs of *L. interrogans* serovar Lai genome had the following consensus vote predicted localizations: 418 cytoplasmic proteins (CP), 332 cytoplasmic membrane proteins (CM), 17 periplasmic proteins (PP), 15 outer membrane proteins (OM), and 15 extracellular/secreted proteins (EX) (Table 2, 3, 4 Additional file 1, 2, 3). The biological functions of most of the localized proteins are already annotated. Only about 9% (68 of 797 ORFs) were proteins annotated as conserved hypothetical or unknown proteins. This shows that the consensus vote approach has a high accuracy of subcellular localization prediction for *L. interrogans*. However, this recall of these methods is unacceptably low, since the localization of the

majority of proteins remains unknown (3930 out of 4727 proteins).

When comparing the concordance or prediction agreement rates between the three prediction methods (excluding proteins with unknown localization by one or two programs), the rates for PSORTb and PA, PSORTb and ProtCompB, and PA and ProtCompB were 70.3%, 80%, and 59.5%, respectively. PSORTb was found to have a strong propensity to assign protein queries to CP and OM proteins, while PA was found to assign preferentially to CM, PP and EX proteins ( $p < 0.001$ , chi-square tests).

#### **Step 2: Homology-based and protein folding recognition predictions for non-consensus vote localizations**

The non-consensus vote OM, EX, and PP proteins were further analyzed for localizations using DBsubloc, and GTD. As presented in Table 5, 6, 99 more proteins (43 out of 83 proteins predicted by two previous methods and 56 out of 617 proteins predicted by one previous method) were additionally identified as putative EX, while 48 proteins (23 out of 59 proteins predicted by two methods,

**Table 3: Putative extracellular proteins (EX) predicted by the consensus vote**

Lai Locus	Copen Locus	Protein annotation
LA3731	LIC10497	Fmh-like protein/hypothetical protein
LA0587	LIC12988	Lactonizing lipase/lipase
LA0872	LIC12760	Microbial collagenase
LA1450	LIC12302	Probable O-sialoglycoprotein endopeptidase
LA2448	LIC10830	Putative outermembrane protein/putative lipoprotein
LA1765	LIC12047	Rhs family protein/cytoplasmic membrane protein
LA4161	LIC13320	Thermolysin/thermolysin precursor
LA4164	LIC13321	Thermolysin/thermolysin homolog precursor
LA2303	LIC11634	3-oxoacyl- [acyl-carrier protein] reductase/CsgA
LA0873	LIC12759	LRR containing protein/cytoplasmic membrane protein
LA2964	LIC11098	LRR containing protein/conserved hypothetical protein
LA3028	LIC11051	LRR containing protein/conserved hypothetical protein
LA3320	LIC10831	LRR containing protein/conserved hypothetical protein
LA3323	LIC10829	LRR containing protein/conserved hypothetical protein
LA0709	LIC12896	Unknown protein/conserved hypothetical protein

Note LRR: Leucine-rich repeat

**Table 4: Putative outer membrane proteins (OM) predicted by the consensus vote**

Lai Locus	Copen Locus	Protein annotation
LA2375	LIC11570	General secretory pathway protein D
LA3149	LIC10964	Hemin receptor/TonB-dependent outer membrane hemin receptor
LB328	LIC20250	Outer membrane protein OmpA/PG-associated CM protein
LA3615	LIC10592	Outer membrane protein OmpA family/PG-associated CM protein
LA1963	LIC11941	Outer membrane protein precursor CzcC/heavy metal efflux pump
LA3927	LIC13135	Outer membrane protein tolC precursor/outer membrane protein
LA1356	LIC12374	Probable TonB-dependent receptor
LA2641	LIC11345	Probable TonB-dependent receptor/ferrichrome-iron receptor
LA3468	LIC10714	Probable TonB-dependent receptor/outer membrane receptor protein
LB191	LIC20151	Putative TonB-dependent outer membrane receptor protein (Hbp A)
LA2510	LIC11458	Conserved hypothetical protein/outer membrane protein, porin superfamily
LA4337	LIC13479	Conserved hypothetical protein/PG-associated CM protein
LA0572	LIC12998	Conserved hypothetical protein/TonB-dependent outer membrane receptor
LA3258	LIC10881	Hypothetical protein/outer membrane protein, TonB dependent
LA2186	LIC11739	Conserved hypothetical protein

Lai locus: *L. interrogans* serovar Lai locusCopen locus: *L. interrogans* serovar Copenhageni locus

and 25 from 980 proteins predicted by one method) were additionally identified as putative OM proteins as shown in Table 7, 8. Moreover, 58 proteins (20 out of 20 proteins predicted by two methods and 38 out of 504 proteins predicted by one method) were additionally predicted as PP proteins (Additional file 1). It is of interest that several protein loci currently annotated as hypothetical proteins without localization information were predicted in EX, OM and PP compartments by the combination method (Tables 3, 4, 5, 6, 7, 8, 9 and Additional file 1). The homology search and structural information from DBSubloc and GTD thus allowed further identification of EX, OM, and PP from the non-consensus vote set, however, 3725 protein localizations remain unknown.

### Step 3: Cytoplasmic (CP) and cytoplasmic membrane proteins (CM) identified by Linear Discriminant Analysis (LDA)

The remaining 3725 proteins with unknown localization after step 2 were further analyzed using an LDA-based classifier we developed to identify CP and CM proteins using the set of CP and CM consensus outputs (418 CP proteins and 332 CM proteins) predicted by all of the three prediction programs (Additional file 2, 3) as a training set (see Materials and Methods). 2272 CP and 481 CM proteins were additionally identified from the 3725 "unknown set" by this approach (Additional file 4, 5). We also found that 66% (1501 out of 2272) of the LDA based predicted CP and 54% (260 out of 481) of the LDA based predicted CM are hypothetical or unknown proteins. In other words, overall 56.3 % (1516 out of 2690) of hypothetical and/or unknown proteins in the whole genome were assigned as CP and 38 % as CM or helix transmembrane proteins.

After the final step in the prediction method, we are able to confidently predict the localization of 3755 (79.4%) Leptospiral proteins. Our combination method thus has a considerably improved recall over the PSORTB and PA methods, approaching that of ProtCompB (Table 1). To test the final prediction accuracy with estimated % agreement and % coverage of our combination method, we then performed the localization prediction of 28 experimentally verified proteins from several studies of Leptospiral outer membrane and extracellular, or cell surface proteins.

### Protein subcellular localization prediction on the experimentally verified leptospiral outer membrane and extracellular proteins

As shown in the Additional file 6, the three prediction programs PSORTb, PA and ProtCompB gave markedly different predictions from one another for 28 experimentally OM and EX. Each of the three prediction programs had weaknesses, either poor agreement (ProtCompB) or low coverage (PSORTb and PA). Our combination approach was much better in the respect and showed good agreement and coverage.

### Discussion

Computational prediction for protein subcellular localization is a key step for genome annotation and development of drug and vaccine target. In this study, we used a combination method to putatively assign CP, CM, PP, OM, and EX proteins. We combined the results from three different algorithms namely PSORTb, PA and ProtCompB into a consensus vote to obtain higher prediction accuracy. The combination approach has previously been used to significantly reduce, or exclude false positive predic-

**Table 5: 43 Putative extracellular proteins (EX) derived from the 2 out of 3 predictions with significant DBSubloc or/and GTD prediction**

Lai Locus	Copen Locus	Protein annotation	SWISS-PROT <sup>a</sup>	PDB Code <sup>b</sup>
LA1027	LIC12632	Sphingomyelinase C precursor (Sph1)/hemolysin	-	Ibix
LA1029	LIC12631	Sphingomyelinase C precursor (Sph2)/hemolysin	-	Ibix
LA4004	LIC13198	Sphingomyelinase C precursor hemolysin (Sph3)/sph-like	-	Ibix
LA3540	LIC10657	Sphingomyelinase C precursor; hemolysin	-	Ibix
LA3050	LIC11040	Hemolytic protein-like protein/hemolysin (sph4)	-	Iaq0
LA3466	LIC10715	Thermolysin	P43133	Ihyt
LA3454	LIC10723	Flagellar hook-associated protein(fliD)	O9KWW7	Iosp
LA3097	LIC11003	Treponemal membrane protein B precursor-like protein/LipL71	P19649	I18w
LA1530	LIC12234	LRR containing protein	O9RBS2	Id0b
LA1324	LIC12401	LRR containing protein/cytoplasmic membrane protein	-	Iogq
LA1354	LIC12375	LRR containing protein/cytoplasmic membrane protein	O9RBS2	Iogq
LA2452	LIC11504	LRR containing protein/cytoplasmic membrane protein	O9RBS2	Iogq
LA2862	LIC11180	LRR containing protein/cytoplasmic membrane protein	O9RBS2	Iogq
LA2966	LIC11097	LRR containing protein/cytoplasmic membrane protein	O9RBS2	Iogq
LA3324	LIC10831	LRR containing protein/conserved hypothetical protein	O9RBS2	Iogq
LA3321	LIC10830	LRR containing protein/putative lipoprotein	O9RBS2	Iogq
LA3322	LIC10830	LRR containing protein/putative lipoprotein	O9RBS2	Iogq
LA0701	LIC12901	LRR containing protein/molybdate metabolism regulator	O9RBS2	Iogq
LA2377	LIC11568	Peptidase, M23/M37/membrane associated peptidase	P24204	Iacc
LA0505	LIC13050	Probable glycosyl hydrolase/conserved hypothetical protein	-	If00
LA3725	LIC10502	Probable phenazine biosynthesis family protein/CM protein	-	Iair
LA3730	LIC10498	Putative lipoprotein	P15921	Irmg
LA1368	LIC12364	Putative outer membrane protein/CagA	P47460	-
LA1759	LIC12050	Putative outer membrane protein/conserved hypothetical protein	O52657	Iczf
LA2443	LIC11507	Putative outer membrane protein/conserved hypothetical protein	O9RBS2	Iogq
LA2447	LIC11505	Putative outer membrane protein/conserved hypothetical protein	O9RBS2	Ij15
LA2450	LIC11505	Putative outer membrane protein/conserved hypothetical protein	O9RBS2	Iogq
LA1915	LIC11990	TPR-repeat-containing proteins/cytoplasmic membrane protein	P80544	Iqqe
LA0043	LIC10038	TPR-repeat-containing proteins/conserved hypothetical protein	O9KO40	Iqqe
LA2773	LIC11246	Conserved hypothetical protein	O06852	I18w
LA3233	LIC10902	Conserved hypothetical protein	O83497	Iqcx
LB001	LIC20001	Conserved hypothetical protein	-	Ieut
LA1499	LIC12259	Conserved hypothetical protein/cytoplasmic membrane protein	P35825	Idab
LA1766	LIC12047	Conserved hypothetical protein/cytoplasmic membrane protein	O07833	Iczf
LA3333	LIC10825	Conserved hypothetical protein/cytoplasmic membrane protein	O07833	Iacc
LA2208	LIC11720	Conserved hypothetical protein/hypothetical protein	-	Ie15
LA3276 <sup>c</sup>	LIC10868	Conserved hypothetical protein/hypothetical protein	P15345	Idab
LA0022	LIC10021	Conserved hypothetical protein/putative lipoprotein	-	Idab
LA3210	LIC10920	Conserved hypothetical protein/putative lipoprotein	-	Irmg
LA3726	LIC10501	Conserved hypothetical protein/putative lipoprotein	O9PJY2	Iacc
LB216	LIC20172	Conserved hypothetical protein/putative lipoprotein	-	Iwxr
LB225	LIC20176	Conserved hypothetical protein/putative lipoprotein	-	Iwxr
LA4135 <sup>d</sup>	LIC13296	hypothetical protein/putative lipoprotein	-	Ikoe

Note LRR: Leucine rich repeat

a: Swiss-Prot ID derived from DBsubloc database

b: PDB code derived from GTD prediction

c: Pfam: PF06739: SBBP (Seven Beta Blade Propeller domain)

d: pfam07588: DUF1554

tions for membrane topology prediction [33], and outer membrane prediction [34]. In our case, the accuracy of consensus vote is very high, since well characterized OM and EX proteins were predicted including lactonizing lipase [35], microbial collagenase [36], O-sialoglycoprotein endopeptidase [37], Rhs family protein [38], CsgA or C factor [39], thermolysin [40], leucine rich repeat pro-

teins (LRR) [41-43], Ton-B dependent outer membrane receptor proteins, OmpA, porin, heavy metal efflux pump, TolC, and general secretory pathway protein D (Table 4).

On the other hand, the recall, or sensitivity of consensus vote prediction is low, especially for EX and OM. The recall for consensus vote is low, because PSORTb and PA

**Table 6: 56 Putative extracellular proteins (EX) derived from the 1 out of 3 predictions with significant DBSubloc or/and GTD prediction**

Lai Locus	Copen locus	Protein annotation	SWISS-PROT <sup>a</sup>	PDB code <sup>b</sup>
LB258	LIC20197	Cysteine protease	-	Ideu
LA0975	LIC12680	Fimh-like protein	-	Ia6c
LA0858	LIC12930	Fimh-like protein/hypothetical protein	-	Idab
LA0492	LIC13060	LipL36 protein	-	Iacc
LA3469	LIC10713	Iron-regulated protein A/LruB/putative lipoprotein	-	Irmg
LA3075	LIC10464	Surface protein Lk90-like protein/Ig-like repeat domain	P35828	Idab
LA3778	LIC10464	Surface protein Lk90-like protein/Ig-like repeat domain	Q52657	Idbg
LA0378	LIC10325	TPR-repeat-containing proteins/hemolysin	Q98KCI	Ia17
LA3138	LIC10973	Transmembrane outer membrane protein L1	-	Iacc
LA1353	LIC12375	LRR containing protein	Q9RBS2	Ij15
LB196	LIC20154	LRR containing protein/lipoprotein	-	Id0b
LA0416 <sup>e</sup>	LIC10365	Putative lipoprotein (LpL effector)	-	Igq8
LA0962 <sup>d</sup>	LIC12690	Putative lipoprotein	-	Ieut/Ikoe
LA1569 <sup>c</sup>	LIC12208	Putative lipoprotein	P15345	Iacc
LA2823 <sup>e</sup>	LIC11207	Putative lipoprotein	-	Igq8
LA3064 <sup>e</sup>	LIC11030	Putative lipoprotein	-	Iczf
LA3848 <sup>c</sup>	LIC13075	Putative lipoprotein	-	Iqiv
LA3867	LIC13086	Putative lipoprotein	-	Icvv
LA1159	LIC12525	Putative outer membrane protein/putative lipoprotein	-	Ics6
LA1905	LIC11996	Putative outer membrane protein/hypothetical protein	-	Ikit
LA1939	LIC11966	Putative outer membrane protein/hypothetical protein	-	Ifio
LA2273	LIC11665	Putative outer membrane protein/hypothetical protein	-	Iair
LA0563 <sup>d</sup>	LIC13006	Hypothetical protein/putative lipoprotein (LenC)	-	Ikoe
LA0695 <sup>d</sup>	LIC12906	Hypothetical protein/putative lipoprotein (LenA/LfhA/Lsa24)	-	Ikoe
LA1433 <sup>d</sup>	LIC12315	Hypothetical protein/putative lipoprotein (LenD)	-	Ikoe
LA3103 <sup>d</sup>	LIC10997	Hypothetical protein (LenB)	-	Ikoe
LA4073 <sup>d</sup>	LIC13248	Hypothetical protein/putative lipoprotein (LenF)	-	Ikoe
LA4324 <sup>d</sup>	LIC13467	Hypothetical protein/conserved hypothetical protein (LenE)	-	Ikoe
LA3370	LIC10793	Conserved hypothetical protein/surface antigen (Lp24)	-	Il0q
LA0965	LIC12676	Conserved hypothetical protein	P25156	Id0b
LA1066	LIC12601	Conserved hypothetical protein	-	Idbg
LA1498	LIC12260	Conserved hypothetical protein	-	Iogq
LA2811	LIC11217	Conserved hypothetical protein	P25146	Iogq
LA3734	LIC10495	Conserved hypothetical protein/CM protein	-	Idab
LA3834 <sup>c</sup>	LIC13066	Conserved hypothetical protein	P15345	Iacc
LA4227	LIC13381	Conserved hypothetical protein	-	Isli
LA0663	LIC12930	Conserved hypothetical protein/hypothetical protein	-	Iacc
LA0423 <sup>c</sup>	LIC10371	Conserved hypothetical protein/putative lipoprotein	P15345	Iqiv
LA1567 <sup>c</sup>	LIC12209	Conserved hypothetical protein/putative lipoprotein	P15345	Iczf
LA1568 <sup>c</sup>	LIC12209	Conserved hypothetical protein/putative lipoprotein	P15345	Iczf/Idbg
LA1691 <sup>c</sup>	LIC12099	Conserved hypothetical protein/putative lipoprotein	-	Iacc
LA3340 <sup>e</sup>	LIC10821	Conserved hypothetical protein/putative lipoprotein	-	Iee6
LA3394 <sup>e</sup>	LIC10774	Conserved hypothetical protein/putative lipoprotein	-	Igq8
LA3501	LIC10686	Conserved hypothetical protein/putative lipoprotein	-	Iair
LA0283 <sup>c</sup>	LIC10239	Hypothetical protein	-	Iair
LA0426 <sup>c</sup>	LIC10373	Hypothetical protein	P56964	Iacc
LA0996 <sup>d</sup>	LIC12668	Hypothetical protein	-	Ikoe
LA1764	LIC12048	Hypothetical protein	-	Iqlg
LA1869	LIC12023	Hypothetical protein	-	Ik14
LA2272	LIC11664	Hypothetical protein	-	Idab
LA3240	LIC10898	Hypothetical protein	-	Irmlg
LA0074	LIC10067	Hypothetical protein/conserved hypothetical protein	-	Idbg
LA1065	LIC12602	Hypothetical protein/conserved hypothetical protein	-	Idab
LA1762	LIC12048	Hypothetical protein/conserved hypothetical protein	-	Iqcx
LA3649	LIC10561	Hypothetical protein/conserved hypothetical protein	-	Iqcx
LA3881	LIC13101	Hypothetical protein/OM with integrin like repeat domains	P35825	Idab

Note LRR: Leucine-rich repeat, a: Swiss-Prot ID derived from DBsubloc database, b: PDB code derived from GTD prediction, c: pfam06739: SBBP (seven bladed beta propeller) repeat d: pfam07588: DUF1554, e: pfam07602: DUF1565

**Table 7: 23 Putative outer membrane proteins (OM) derived the 2 out of 3 predictions with significant DBSubloc or/and GTD prediction**

Lai locus	LIC locus	Protein annotation	SWISS-PROT <sup>a</sup>	PDB code <sup>b</sup>
LA3471	LIC10711	Iron-regulated protein A/cytoplasmic membrane protein	<u>P12608</u>	<u>1i5p</u>
LA1161	LIC12524	Long-chain fatty acid transport protein/fatty acid transport protein	-	<u>1kmo</u>
LA1100	LIC12575	Outer membrane efflux protein/cytoplasmic membrane protein	-	<u>1ek9</u>
LA1445	LIC12307	Outer membrane efflux protein/OM- TolC superfamily	<u>P50468</u>	<u>1ek9</u>
LA3685	LIC10537	Outer membrane protein/PG- associated periplasmic protein	<u>P38369</u>	<u>1rlm</u>
LA0056	LIC10050	Outer membrane protein OmpA family/PG-associated CM protein	<u>Q05146</u>	<u>1rlm</u>
LA2318	LIC11623	Predicted outer membrane protein/outer membrane protein	-	<u>1a0t</u>
LA1968	LIC11935	Putative outer membrane protein/conserved hypothetical protein	-	<u>1a0t</u>
LA2444	LIC11506	Putative outer membrane protein/outer membrane protein	-	<u>1fep</u>
LB110	LIC20087	Putative outer membrane protein/outer membrane protein	-	<u>1uyu</u>
LA2242	LIC11694	TonB-dependent outer membrane receptor	<u>P46359</u>	<u>1fep</u>
LA3242	LIC10896	TonB-dependent outer membrane receptor	<u>P37409</u>	<u>1kmo</u>
LA0465	LIC10405	TPR-repeat-containing proteins/conserved hypothetical	<u>P58937</u>	-
LA3675	LIC10544	Hypothetical protein/outer membrane protein	-	<u>1a0t</u>
LA2063	LIC11851	Conserved hypothetical protein/cytoplasmic membrane protein	-	<u>1by5</u>
LA3102	LIC10998	Conserved hypothetical protein	<u>P76115</u>	<u>1nqe</u>
LA3675	LIC10544	Hypothetical protein/outer membrane protein	-	<u>1a0t</u>
LA2168	-	Hypothetical protein	<u>P43153</u>	<u>1a0t</u>
LA3809	LIC10439	Hypothetical protein	-	<u>1a0t</u>
LA1501	LIC12258	Hypothetical protein	-	<u>2mpr</u>
LA3552	LIC10647	Hypothetical protein/conserved hypothetical protein	-	<u>1kmo</u>
LA2818	LIC11211	Hypothetical protein/conserved hypothetical protein	-	<u>2mpr</u>
LA4059	LIC13238	Hypothetical protein/conserved hypothetical protein	-	<u>1by5</u>
LB279	LIC20214	Hypothetical protein/conserved hypothetical protein	-	<u>1kmo</u>

Note a: Swiss-Prot ID derived from DBsubloc database, b: PDB code derived from GTD prediction

**Table 8: 25 Putative outer membrane proteins (OM) derived from the 1 out of 3 predictions with significant DBSubloc and/or GTD prediction**

Lai locus	Copen Locus	Protein annotation	SWISS-PROT <sup>a</sup>	PDB Code <sup>b</sup>
LA0616	LIC12966	LipL41/Outer membrane lipoprotein lipL41	-	<u>1a17</u>
LA2295	LIC11643	LipL45 protein	<u>P02977</u>	<u>1l8w</u>
LA0957	LIC12693	Outer membrane efflux protein/conserved hypothetical protein	<u>P24145</u>	<u>1ek9</u>
LA0581	LIC12990	Outer membrane efflux protein/conserved hypothetical protein	<u>Q9ZHD2</u>	<u>1ek9</u>
LA3733	LIC10496	Outer membrane efflux protein/conserved hypothetical protein	-	<u>1ek9</u>
LA0301	LIC10258	Outer membrane protein OmpA family/hypothetical protein	<u>Q926C3</u>	<u>1rlm</u>
LA0222	LIC10191	Outer membrane protein OmpA family/PG-associated CM protein	<u>P22263</u>	<u>1rlm</u>
LA1192	LIC12499	Putative outer membrane protein	-	<u>1fep</u>
LA1404	LIC12337	Putative outer membrane protein	-	<u>2mpr</u>
LA1931	LIC11975	Putative outer membrane protein/outer membrane protein	-	<u>2mpr</u>
LA1987	LIC11918	Putative outer membrane protein/conserved hypothetical protein	-	<u>1osp</u>
LB199	LIC20157	Putative outer membrane protein/conserved hypothetical protein	-	<u>1fep</u>
LA1030	LIC12630	TPR-repeat-containing proteins/hypothetical protein	<u>P58937</u>	<u>1a17</u>
LA0568	LIC13002	Conserved hypothetical protein	-	<u>1kmo</u>
LA1510	LIC12252	Conserved hypothetical protein	-	<u>2mpr</u>
LA0835	LIC12791	Hypothetical protein/conserved hypothetical protein	-	<u>1fnf</u>
LA2746	LIC11268	Hypothetical protein/conserved hypothetical protein	-	<u>2mpr</u>
LA2940	LIC11121	Hypothetical protein/conserved hypothetical protein	-	<u>2mpr</u>
LA2976	LIC11086	Hypothetical protein/conserved hypothetical protein	-	<u>2mpr</u>
LA3870	LIC13089	Hypothetical protein/conserved hypothetical protein	-	<u>2mpr</u>
LA4272	LIC13418	Hypothetical protein/conserved hypothetical protein	-	<u>2mpr</u>
LA4335	LIC13477	Hypothetical protein/conserved hypothetical protein	-	<u>1kmo</u>
LA0706	LIC12898	Unknown protein	<u>P38370</u>	<u>1fep</u>
LA1507	LIC12254	Unknown protein/outer membrane protein	-	<u>1a0t</u>
LA3853	LIC13078	Unknown protein/conserved hypothetical protein	-	<u>1bxw</u>

Note a: Swiss-Prot ID derived from DBsubloc database, b: PDB code derived from GTD prediction

**Table 9: Protein subcellular localizations of *L. interrogans* predicted by PSORTb, PA, ProtCompB and the combination prediction**

Localization	Subcellular localization prediction			
	PSORTb	PA	ProtCompB	Combination prediction
Cytoplasm (CP)	1125	921	2013	2690
Cytoplasmic membrane (CM)	606	715	1726*	813
Outer membrane (OM)	112	28	-	63
Periplasmic (PP)	30	86	478	75
Extracellular (EX)	29	326	510	114
Unknown	2825	2652	-	972

\* Note that in ProtCompB prediction in this version, CM and OM were predicted as membrane proteins.

programs are known to have limitations for some proteins. PSORTb requires a training set from a limited number of experimentally-determined proteins, while PA has a disadvantage in that query proteins have to share similarity to known proteins in the Swiss-Prot database [44]. Among high-throughput computational predictions for protein subcellular localization, PSORTb has been reported as the prediction tool that achieves the highest overall accuracy, followed closely by PA [22].

To overcome the limitations in PSORTb, PA and ProtCompB, the predictions for proteins predicted by only one or two out of the three prediction methods (the non consensus vote) were refined by homology-based search using the DBSubloc database and structural annotation in GTD. This allowed us to identify protein localizations with greater confidence. The advantage of GTD is that protein folding recognition or threading methods can determine pairs of proteins that have no obvious similarities in sequence, but have similar folds. It was previously suggested this approach should be carried out to increase prediction sensitivity for specific protein localization [22,45,46]. To our knowledge, this study is the first to employ GTD information to infer leptospiral protein localizations.

Structure-based information from GTD prediction revealed that the majority of the 99 EX predictions were proteins that may be secreted by the type III or the type V (autotransport) system. These proteins are shown in Table 5, 6 with their corresponding PDB code. Many of the putative EX proteins that are annotated as leucine rich repeat (LRR) containing proteins share sequence similarity to PopC protein (Q9RBS2), which is secreted through the hrp-secretion apparatus or the type III secretion pathway of *Ralstonia solanacearum* [41]. Structurally related well-characterized extracellular LRR proteins in other species include YopM (PDB code 1jl5), a *Yersinia pestis* cytotoxin [43], internalin B [47], a virulence factor of *Listeria monocytogenes* (PDB code 1d0b) and polygalacturonase inhib-

iting protein (PDB code 1ogq), a secreted protein involved in plant defense [48].

It is of interest to note that several *L. interrogans* proteins are contained within the LRR and TPR (Tetratricopeptide repeat) protein families, but predicted sub-cellular localization is not necessarily conserved among all members within each family (Table 3, 5, 6, 7, 8, 9 and Table in additional file 4). The majority of LRR proteins were predicted to be EX localized, while TPR proteins were predicted in all compartments except PP. This finding is consistent with the multiple functions of TPR homologues from more distantly related species in different sub-cellular milieux, including signal transduction, chaperone activity, cell-cycle, transcription, and protein transport [49,50].

Out of 48 non-consensus vote of predicted OM, 24 were proteins annotated as outer membrane or putative outer membrane proteins, while of the remainder were proteins annotated as conserved hypothetical proteins. The structural information derived from the GTD prediction of the conserved or hypothetical proteins that were predicted as putative OM were the same as that of the annotated outer membrane proteins. As shown in Table 7, 8, it can be observed that 24 hypothetical proteins can now be annotated as putative OM.

Although it is clear that the consensus vote combined with DB and GTD prediction can give robust prediction for EX, OM and PP, there are many proteins with either CP or CM localization remaining. Using our combination approach, we found that about 17% of genes encode putative CM proteins in *L. interrogans* serovar Lai genome, which is of similar proportion to the 20% – 30% CM proteins in other bacterial species [25,51]. From our subcellular location prediction we identified 63 OM and 114 EX proteins as potential vaccine candidates. On the other hand, it is possible to exclude 813 CM and 75 PP predicted proteins as vaccine candidates, on the basis of their localization.

We compared our predictions with the previously published works. We found that 10 of 16 membrane proteins predicted by Gamberini *et al.* 2006, including four also demonstrated to be immunogenic among 8 pathogenic serovars in that study, were also predicted by our method as membrane proteins (2 EX, 1OM, 1PP and 6 CM) [18]. We examined the localizations of the 145 putative lipoproteins reported by Setubal *et al.* [19], and found 29 EX, 2 OM, 7 PP and 26 CM proteins among 125 probable lipoproteins, and 1 PP and 3 CM among 21 possible lipoproteins. The localizations of 63 putative lipoproteins could not be identified, which included proteins containing signal peptidase II recognition sites and proteins lacking sequence and/or structural homology to known membrane proteins (see Additional file 7). Spirochaetal lipoproteins are found in four subcellular compartments: the periplasmic leaflet of the cytoplasmic membrane, the periplasmic outer leaflet of the outer membrane, or beyond the outer membrane into the environment as extracellular proteins [52]. Therefore, 15 of the 145 putative lipoproteins identified as CP by our method are unlikely to be lipoproteins because of their localization. These false positive lipoproteins include UDP-glucose 6-dehydrogenase, cell-division protein, regulator of chromosome condensation RCC1 family, and 3-oxoacyl[acyl-carrier protein] reductase. The frequency of falsely-identified lipoproteins just exceeds the reported 1% false positive rate for the SpLip program [52]. Our results can be considered as complementary to those reported by Setubal *et al.* [52], and increase the accuracy of lipoprotein prediction.

We also compared our predictions with the 226 leptospiral surface exposed protein predictions (extracellular, outer membrane, periplasmic, inner (cytoplasmic) membrane by their localization definition) reported by Yang *et al.* [20] and found a concordance of 38.5 % (87/226) (see Additional file 8). We think the discrepancies arise from false assignments generated by the prediction algorithms used, which can be identified by comparison with proteins for which there are reliable experimental data of localization (see Additional file 6) [2-14,53-57]. Our predictions have a higher coverage and agreement with the experimentally tested *L. interrogans* protein set than the study by Yang *et al.* [20], suggesting that our prediction method may be of greater overall utility for genome annotation of membrane proteins. After manual inspection of predicted localizations, we found further examples of possible false assignments. The greatest discrepancy was found for 42 proteins identified as CM by our method, but OM by Yang *et al.* Some proteins among this group have homologues in other species for which there is experimental evidence of CM location, including methyl-accepting chemotaxis protein mcpB [58], aero-

taxis sensor receptor [59], and penicillin-binding protein [60].

It was found that several loci without localization annotation were assigned by the combination prediction method. Therefore, we propose that the annotations with respect to subcellular localization for these loci can be tentatively revised. Among this group of proteins, we noted additional similarities to known protein families. One prominent group with the SBBP domain (seven beta blade propeller proteins, Pfam PF06739) contain 9 hypothetical proteins: LA0283 (LIC10239), LA0423 (LIC10371), LA0426 (LIC10373), LA1567 (LIC12209), LA1568(12209), LA1569 (LIC12208), LA1691 (LIC12099), LA3276 (LIC10868), LA3834 (LIC13066). Three loci annotated as hypothetical proteins or lipoproteins, namely LA0996 (LIC12668), LA0962 (LIC12690), and LIC13296 (LA4135), were predicted as EX localized (shown in Table 5, 6), and may belong to the Len (leptospiral endostatin-like lipoproteins) family, based on conservation of DUF1554 domain (pfam PF07588) and structural similarity to mammalian endostatin-like protein (PDB 1koe). These proteins act as adhesion proteins and bind to host extracellular matrix (ECM) [53,57] or human factor H [56]. (Table 5, 6 and Table in the Additional file 6). Furthermore, three loci LIC11207 (LA2823), LIC10821 (LA3340) and LIC10774 (LA3394) and LIC10365 (LA0416), previously described to have similarity with the leptospiral effector protein [54] were identified as putative EX proteins in agreement with their proposed immunomodulator function.

Our combination prediction method has high agreement and coverage of experimentally verified OM and EX proteins (see Additional file 6). On the other hand, experimental localization studies are limited by insufficient sensitivity to detect low abundance proteins and cross contamination of cellular compartments during sample purification, as discussed previously by Rey *et al.* [21]. It is of note that several predicted PP proteins in this work e.g. FlaB1 periplasmic flagellin (LA2017/LIC11890) have previously been identified as possible PP contaminants in experimental studies of OMV proteins [13,20]; hence our prediction method may help in correct interpretation of future experimental verification studies, thus leading to better predictions in uncharacterized genomes. However, it should be emphasized that no automatic prediction can be accurate without experimental verification.

## Conclusion

In this study, we have demonstrated that the specificity and sensitivity of protein subcellular localization prediction can be improved by incorporation of multiple predictive methods and structural information. By this approach, localizations can be assigned to previously

hypothetical *L. interrogans* proteins. We think this approach is applicable for subcellular localization predictions in other prokaryote proteomes, with the caveat that some predictions are robust than others, i.e. CP and CM better than OM, EX or PP.

## Materials and Methods

### Data sets

Amino acid sequence queries were 4,727 proteins of *Leptospira interrogans* serovar Lai genome (chromosome I: NC\_004342, chromosome II: NC\_004343) [15] and 3,728 protein ORFs of *Leptospira interrogans* serovar Copenhageni strain (Fiocuz L1-I30) [accession number AEO16823 (chromosome I) and AEO16824 (chromosome II) [17] obtained from GenBank. Two datasets of proteins with known subcellular localization were used. One was an experimentally confirmed data set containing 278 CP and 309 CM of Gram-negative bacteria described by Gardy *et al.* 2003 [28] and used for validation of the LDA based classifier's performance. Another one was a 299 protein-data set containing 145 CP, 69 CM proteins, 29 PP, 38 OM and 18 EX which was the testing data previously used to evaluate various protein localization predictions in Gardy and Brinkman [22].

### Computational Data sets mputational prediction tools for in silico protein localization

Several publicly available programs were used in combination of predictions. Protein subcellular localization for Gram-negative bacteria was carried out using PSORTb [27,28], Proteome analysis (PA) [29], and ProtCompB [30]. Feature based predictions for signal peptide sequence and  $\beta$  helix transmembrane proteins were identified using SignalP [23] and TMHMM [24,25] respectively.

### Homology based searching and structural annotation

Homology search for subcellular localization information was carried out using BLAST search against DBSubloc, a localization specific protein database [31]. A protein folding recognition method for structural information used to predict the fold of protein sequence with distant homology to known structure was performed using homology search against GTD (the Genomic Threading Database) [32].

### Prediction strategy (as shown in Figure 1)

#### Step 1. Consensus votes prediction

We reasoned that more accurate protein subcellular localization predictions can be gained from the consensus of methods. All leptospiral protein queries were analyzed using three subcellular localization prediction tools for Gram-negative bacteria, namely PSORTb, Proteome analysis (PA), and ProtCompB for cytoplasm (CP), cytoplasmic membrane (CM), periplasmic (PP), outer membrane

(OM) and extracellular proteins (EX). Note that in this version ProtCompB prediction, CM and OM are not distinguished so both proteins are predicted as membrane proteins. The consensus prediction for each sequence was calculated using a simple majority vote type procedure. If all 3 methods agree for localization, it is assigned as a "consensus vote". The remaining results (1 or 2 out of 3 predicted) were assigned as "non-consensus vote". The CP and CM proteins assigned in this step were used as a training set for the development of LDA based classifier for CP and CM in the next step.

#### Step 2. Homology-based and protein folding recognition prediction

Homology based and structural information can also be used to infer the potential localization site of query proteins [22,45,46]. Therefore, the remaining query proteins assigned as non-consensus vote results of PP, OM and EX were further analyzed for sequence and structure homology. Since subcellular localization is an evolutionarily conserved trait, if a protein query is homologous to a known protein with the same localization, the localization was assigned. The protein query sequences were compared to proteins in DBSubloc database at  $E$ -value  $\leq 10^{-3}$  using BLAST search. Structure annotation of these queries was also performed using GTD prediction. The query proteins sequences were assigned to structures (shown as PDB code) with the high level of probability prediction (certain and high) for these protein queries. In this study, the confidence range based on p-value of measuring the reliability of the structure annotation as certain ( $0 \leq p < 0.01\%$ ) and high ( $0.01\% \leq p < 0.1\%$ ) were considered as a statistically significant structure annotation.

#### Step 3. Identification of putative CP and CM using the LDA based classifier

A number of putative CP and CM identified as non-consensus vote results was further analyzed by SignalP and TMHMM. The feature attributors derived from SignalP and TMHMM predictions were then integrated and analyzed using the LDA based classifier. Proteins classified with probabilities  $\geq 0.9$  to be CP or CM proteins were taken as significant. The remaining queries that could not be identified in this step were classified as "unknown" results.

### LDA based Classifier for CP and CM

We developed a specific classifier using the training set driven from the consensus vote prediction of leptospiral CP and CM proteins to increase the accuracy of prediction. In the classification-based prediction, our classifier was built on an LDA algorithm analyzing the value of multiple character vectors of SignalP-NN, SignalP-HMM and TMHMM prediction results of the set of training sequences. The accuracy of the LDA based classifier was investigated using leave-one out cross validation. We used

experimentally determined or known CP and CM proteins of Gram-negative bacteria previously performed in the evaluation of PSORTb as a test dataset for validation of the LDA based classifier's performance [27]. Overall, the accuracy of LDA based classifier achieved 94.96%.

#### Authors' contributions

WV and SI participated in designed the research project. SI and EP carried out the computational analysis and developed LDA-based classifier. WV analyzed and interpreted the result, drafted and produced the manuscript. PP provided the further insights for refining the manuscript. All authors read and approved the final manuscript.

#### Additional material

##### Additional file 1

*Putative PP proteins in L. interrogans serovar Lai genome.* This table lists the Lai locus and protein annotation of (A) 17 predicted PP derived from the consensus vote prediction (B) 20 predicted PP derived from 2 out of 3 predictions with significant DBsubloc and/or GTD predictions, (C) 38 predicted PP derived from 1 out of 3 predictions with significant DBsubloc and/or GTD predictions.

[Click here for file](#)

[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-181-S1.xls>]

##### Additional file 2

*Putative CP proteins predicted by the consensus vote prediction in L. interrogans serovar Lai genome.* This table lists the Lai locus and protein annotation of 418 predicted CP proteins derived from consensus vote and used as the training set for the development of the LDA based classifier.

[Click here for file](#)

[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-181-S2.xls>]

##### Additional file 3

*Putative CM proteins predicted by the consensus vote prediction in L. interrogans serovar Lai genome.* This table lists the Lai locus and protein annotation of 332 predicted CM proteins derived from consensus vote and used as the training set for the development of the LDA based classifier.

[Click here for file](#)

[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-181-S3.xls>]

##### Additional file 4

*Putative CP proteins predicted by LDA based classifier of L. interrogans serovar Lai genome.* This table lists the Lai locus and protein annotation of 2272 predicted CP proteins predicted by LDA based classifier

[Click here for file](#)

[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-181-S4.xls>]

##### Additional file 5

*Putative CM proteins predicted by LDA based classifier of L. interrogans serovar Lai genome.* This table lists the Lai locus and protein annotation of 481 predicted CM proteins predicted by LDA based classifier.

[Click here for file](#)

[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-181-S5.xls>]

##### Additional file 6

*Subcellular localizations of 28 experimentally studied OM and EX proteins of L. interrogans serovar Lai.* This table lists the protein name, L. interrogans serovar Lai and copenhageni locus, experimental localization, subcellular localization prediction using PSORTb, ProtCompB, PA, and the combination prediction of 28 experimentally studied OM and EX proteins.

[Click here for file](#)

[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-181-S6.xls>]

##### Additional file 7

*The result of subcellular localization of putative lipoproteins using the combination method.* This table lists the Lai locus tag and protein annotation of 125 probable lipoproteins and 21 possible lipoproteins predicted by SpiLip programs [19] and the subcellular localization of these lipoproteins predicted by the combination method.

[Click here for file](#)

[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-181-S7.xls>]

##### Additional file 8

*Subcellular localization of vaccine candidate using the combination method..* This table lists the Lai locus tag and protein annotation of 226 vaccine candidate predicted by Yang et al. [20] and the subcellular localization of these vaccine candidates predicted by the combination method.

[Click here for file](#)

[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-181-S8.xls>]

#### Acknowledgements

We greatly thank Philip Shaw, Sastra Chaotheing and Duangdoa Wichadakul for their helpful critical reading and commend of the manuscript. This work was supported by the grant from the National Center for Genetic Engineering and Biotechnology, Thailand.

#### References

- Bharti AR, Nally JE, Ricardi JN, Matthias MA, Diaz MM, Lovett MA, Levett PN, Gilman RH, Willig MR, Gotuzzo E, Vinetz JM: **Leptospirosis: a zoonotic disease of global importance.** *Lancet Infect Dis* 2003, **3(12)**:757-771.
- Haake DA, Champion CI, Martinich C, Shang ES, Blanco DR, Miller JN, Lovett MA: **Molecular cloning and sequence analysis of the gene encoding OmpL1, a transmembrane outer membrane protein of pathogenic Leptospira spp.** *J Bacteriol* 1993, **175(13)**:4225-4234.
- Shang ES, Summers TA, Haake DA: **Molecular cloning and sequence analysis of the gene encoding LipL41, a surface-exposed lipoprotein of pathogenic Leptospira species.** *Infect Immun* 1996, **64(6)**:2322-2330.
- Haake DA, Martinich C, Summers TA, Shang ES, Pruett JD, McCoy AM, Mazel MK, Bolin CA: **Characterization of leptospiral outer membrane lipoprotein LipL36: downregulation associated**

- with late-log-phase growth and mammalian infection.** *Infect Immun* 1998, **66**(4):1579-1587.
5. Haake DA, Chao G, Zuerner RL, Barnett JK, Barnett D, Mazel M, Matsunaga J, Levett PN, Bolin CA: **The leptospiral major outer membrane protein LipL32 is a lipoprotein expressed during mammalian infection.** *Infect Immun* 2000, **68**(4):2276-2285.
  6. Lee SH, Kim KA, Park YG, Seong IW, Kim MJ, Lee YJ: **Identification and partial characterization of a novel hemolysin from Leptospira interrogans serovar Iai.** *Gene* 2000, **254**(1-2):19-28.
  7. Cullen PA, Cordwell SJ, Bulach DM, Haake DA, Adler B: **Global analysis of outer membrane proteins from Leptospira interrogans serovar Lai.** *Infect Immun* 2002, **70**(5):2311-2318.
  8. Haake DA, Matsunaga J: **Characterization of the leptospiral outer membrane and description of three novel leptospiral membrane proteins.** *Infect Immun* 2002, **70**(9):4936-4945.
  9. Cullen PA, Haake DA, Bulach DM, Zuerner RL, Adler B: **LipL21 is a novel surface-exposed lipoprotein of pathogenic Leptospira species.** *Infect Immun* 2003, **71**(5):2414-2421.
  10. Koizumi N, Watanabe H: **Molecular cloning and characterization of a novel leptospiral lipoprotein with OmpA domain.** *FEMS Microbiol Lett* 2003, **226**(2):215-219.
  11. Matsunaga J, Barocchi MA, Croda J, Young TA, Sanchez Y, Siqueira I, Bolin CA, Reis MG, Riley LW, Haake DA, Ko AI: **Pathogenic Leptospira species express surface-exposed proteins belonging to the bacterial immunoglobulin superfamily.** *Mol Microbiol* 2003, **49**(4):929-945.
  12. Zhang YX, Geng Y, Bi B, He JY, Wu CF, Guo XK, Zhao GP: **Identification and classification of all potential hemolysins encoding genes and their products from Leptospira interrogans serogroup Icterohaemorrhagiae serovar Lai.** *Acta Pharmacol Sin* 2005, **26**(4):453-461.
  13. Nally JE, Whitelegge JP, Aguilera R, Pereira MM, Blanco DR, Lovett MA: **Purification and proteomic analysis of outer membrane vesicles from a clinical isolate of Leptospira interrogans serovar Copenhageni.** *Proteomics* 2005, **5**(1):144-152.
  14. Asuthkar S, Velineiri S, Stadlmann J, Altmann F, Sritharan M: **Expression and characterization of an iron-regulated hemin-binding protein, HbpA, from Leptospira interrogans serovar Lai.** *Infect Immun* 2007, **75**(9):4582-4591.
  15. Ren SX, Fu G, Jiang XG, Zeng R, Miao YG, Xu H, Zhang YX, Xiong H, Lu G, Lu LF, Jiang HQ, Jia J, Tu YF, Jiang JX, Gu WY, Zhang YQ, Cai Z, Sheng HH, Yin HF, Zhang Y, Zhu GF, Wan M, Huang HL, Qian Z, Wang SY, Ma W, Yao ZJ, Shen Y, Qiang BQ, Xia QC, Guo XK, Danchin A, Saint Girons I, Somerville RL, Wen YM, Shi MH, Chen Z, Xu JG, Zhao GP: **Unique physiological and pathogenic features of Leptospira interrogans revealed by whole-genome sequencing.** *Nature* 2003, **422**(6934):888-893.
  16. Nascimento AL, Ko AI, Martins EA, Monteiro-Vitorelo CB, Ho PL, Haake DA, Verjovski-Almeida S, Hartskeerl RA, Marques MV, Oliveira MC, Menck CF, Leite LC, Carrer H, Coutinho LL, Degraeve WM, Dellagostin OA, El-Dorry H, Ferro ES, Ferro MI, Furlan LR, Gamberini M, Giglioti EA, Goes-Neto A, Goldman GH, Goldman MH, Harakava R, Jeronimo SM, Junqueira-de-Azevedo IL, Kimura ET, Kuramae EE, Lemos EG, Lemos MV, Marino CL, Nunes LR, de Oliveira RC, Pereira GG, Reis MS, Schriefer A, Siqueira WJ, Sommer P, Tsai SM, Simpson AJ, Ferro JA, Camargo LE, Kitajima JP, Setubal JC, Van Sluys MA: **Comparative genomics of two Leptospira interrogans serovars reveals novel insights into physiology and pathogenesis.** *J Bacteriol* 2004, **186**(7):2164-2172.
  17. Nascimento AL, Verjovski-Almeida S, Van Sluys MA, Monteiro-Vitorelo CB, Camargo LE, Digiampietri LA, Harstkeerl RA, Ho PL, Marques MV, Oliveira MC, Setubal JC, Haake DA, Martins EA: **Genome features of Leptospira interrogans serovar Copenhageni.** *Braz J Med Biol Res* 2004, **37**(4):459-477.
  18. Gamberini M, Gomez RM, Atzinger MV, Martins EA, Vasconcellos SA, Romero EC, Leite LC, Ho PL, Nascimento AL: **Whole-genome analysis of Leptospira interrogans to identify potential vaccine candidates against leptospirosis.** *FEMS Microbiol Lett* 2005, **244**(2):305-313.
  19. Setubal JC, Reis M, Matsunaga J, Haake DA: **Lipoprotein computational prediction in spirochaetal genomes.** *Microbiology* 2006, **152**(Pt 1):113-121.
  20. Yang HL, Zhu YZ, Qin JH, He P, Jiang XC, Zhao GP, Guo XK: **In silico and microarray-based genomic approaches to identifying potential vaccine candidates against Leptospira interrogans.** *BMC Genomics* 2006, **7**:293.
  21. Rey S, Gardy JL, Brinkman FS: **Assessing the precision of high-throughput computational and laboratory approaches for the genome-wide identification of protein subcellular localization in bacteria.** *BMC Genomics* 2005, **6**:162.
  22. Gardy JL, Brinkman FS: **Methods for predicting bacterial protein subcellular localization.** *Nat Rev Microbiol* 2006, **4**(10):741-751.
  23. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0.** *J Mol Biol* 2004, **340**(4):783-795.
  24. Emanuelsson O, Brunak S, von Heijne G, Nielsen H: **Locating proteins in the cell using TargetP, SignalP and related tools.** *Nat Protoc* 2007, **2**(4):953-971.
  25. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *J Mol Biol* 2001, **305**(3):567-580.
  26. Kall L, Krogh A, Sonnhammer EL: **A combined transmembrane topology and signal peptide prediction method.** *J Mol Biol* 2004, **338**(5):1027-1036.
  27. Gardy JL, Spencer C, Wang K, Ester M, Tusnady GE, Simon I, Hua S, deFays K, Lambert C, Nakai K, Brinkman FS: **PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria.** *Nucleic Acids Res* 2003, **31**(13):3613-3617.
  28. Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, Brinkman FS: **PSORTb v2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis.** *Bioinformatics* 2005, **21**(5):617-623.
  29. Lu Z, Szafron D, Greiner R, Lu P, Wishart DS, Poulin B, Anvik J, Macdonell C, Eisner R: **Predicting subcellular localization of proteins using machine-learned classifiers.** *Bioinformatics* 2004, **20**(4):547-556.
  30. **ProtCompB - Prediction sub-cellular protein localization** [<http://linux1.softberry.com/berry.php?topic=protcom&pan&group=programs&subgroup=proloc>]
  31. Guo T, Hua S, Ji X, Sun Z: **DBSubLoc: database of protein subcellular localization.** *Nucleic Acids Res* 2004, **32**(Database issue):D122-4.
  32. McGuffin LJ, Street SA, Bryson K, Sorensen SA, Jones DT: **The Genomic Threading Database: a comprehensive resource for structural annotations of the genomes from key organisms.** *Nucleic Acids Res* 2004, **32**(Database issue):D196-9.
  33. Moller S, Croning MD, Apweiler R: **Evaluation of methods for the prediction of membrane spanning regions.** *Bioinformatics* 2001, **17**(7):646-653.
  34. Bagos PG, Liakopoulos TD, Hamodrakas SJ: **Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method.** *BMC Bioinformatics* 2005, **6**:7.
  35. Ihara F, Kageyama Y, Hirata M, Nihira T, Yamada Y: **Purification, characterization, and molecular cloning of lactonizing lipase from Pseudomonas species.** *J Biol Chem* 1991, **266**(27):18135-18140.
  36. Matsushita O, Yoshihara K, Katayama S, Minami J, Okabe A: **Purification and characterization of Clostridium perfringens 120-kilodalton collagenase and nucleotide sequence of the corresponding gene.** *J Bacteriol* 1994, **176**(1):149-156.
  37. Abdullah KM, Lo RY, Mellors A: **Cloning, nucleotide sequence, and expression of the Pasteurella haemolytica A1 glycoprotease gene.** *J Bacteriol* 1991, **173**(18):5597-5603.
  38. Hill CW, Sandt CH, Vlazny DA: **Rhs elements of Escherichia coli: a family of genetic composites each encoding a large mosaic protein.** *Mol Microbiol* 1994, **12**(6):865-871.
  39. Tukel C, Raffatellu M, Humphries AD, Wilson RP, Andrews-Polymeris HL, Gull T, Figueiredo JF, Wong MH, Michelsen KS, Akcelik M, Adams LG, Baumler AJ: **CsgA is a pathogen-associated molecular pattern of Salmonella enterica serotype Typhimurium that is recognized by Toll-like receptor 2.** *Mol Microbiol* 2005, **58**(1):289-304.
  40. Tran L, Wu XC, Wong SL: **Cloning and expression of a novel protease gene encoding an extracellular neutral protease from Bacillus subtilis.** *J Bacteriol* 1991, **173**(20):6364-6372.
  41. Gueneron M, Timmers AC, Boucher C, Arlat M: **Two novel proteins, PopB, which has functional nuclear localization signals, and PopC, which has a large leucine-rich repeat domain, are secreted through the hrp-secretion apparatus of Ralstonia solanacearum.** *Mol Microbiol* 2000, **36**(2):261-277.

42. Ikegami A, Honma K, Sharma A, Kuramitsu HK: **Multiple functions of the leucine-rich repeat protein LrrA of Treponema denticola.** *Infect Immun* 2004, **72**(8):4619-4627.
43. Evdokimov AG, Anderson DE, Routzahn KM, Waugh DS: **Unusual molecular architecture of the Yersinia pestis cytotoxin YopM: a leucine-rich repeat protein with the shortest repeating unit.** *J Mol Biol* 2001, **312**(4):807-821.
44. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbaut S, Schneider M: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31**(1):365-370.
45. Nair R, Rost B: **Sequence conserved for subcellular localization.** *Protein Sci* 2002, **11**(12):2836-2847.
46. Nair R, Rost B: **Better prediction of sub-cellular localization by combining evolutionary and structural information.** *Proteins* 2003, **53**(4):917-930.
47. Bierne H, Sabet C, Personnic N, Cossart P: **Internalins: a complex family of leucine-rich repeat-containing proteins in Listeria monocytogenes.** *Microbes Infect* 2007, **9**(10):1156-1166.
48. Di Matteo A, Federici L, Mattei B, Salvi G, Johnson KA, Savino C, De Lorenzo G, Tsernoglou D, Cervone F: **The crystal structure of polygalacturonase-inhibiting protein (PGIP), a leucine-rich repeat protein involved in plant defense.** *Proc Natl Acad Sci U S A* 2003, **100**(17):10124-10128.
49. D'Andrea LD, Regan L: **TPR proteins: the versatile helix.** *Trends Biochem Sci* 2003, **28**(12):655-662.
50. Blatch GL, Lassle M: **The tetratricopeptide repeat: a structural motif mediating protein-protein interactions.** *Bioessays* 1999, **21**(11):932-939.
51. Wallin E, von Heijne G: **Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms.** *Protein Sci* 1998, **7**(4):1029-1038.
52. Haake DA: **Spirochaetal lipoproteins and pathogenesis.** *Microbiology* 2000, **146** ( Pt 7):1491-1504.
53. Stevenson B, Choy HA, Pinne M, Rotondi ML, Miller MC, Demoll E, Kraiczy P, Cooley AE, Creamer TP, Suchard MA, Brissette CA, Verma A, Haake DA: **Leptospira interrogans Endostatin-Like Outer Membrane Proteins Bind Host Fibronectin, Laminin and Regulators of Complement.** *PLoS ONE* 2007, **2**(11):e1188.
54. Vieira ML, D'Atri LP, Schattner M, Habarta AM, Barbosa AS, de Morais ZM, Vasconcellos SA, Abreu PA, Gomez RM, Nascimento AL: **A novel leptospiral protein increases ICAM-1 and E-selectin expression in human umbilical vein endothelial cells.** *FEMS Microbiol Lett* 2007, **276**(2):172-180.
55. Neves FO, Abreu PA, Vasconcellos SA, de Morais ZM, Romero EC, Nascimento AL: **Identification of a novel potential antigen for early-phase serodiagnosis of leptospirosis.** *Arch Microbiol* 2007, **188**(5):523-532.
56. Barbosa AS, Abreu PA, Neves FO, Atzinger MV, Watanabe MM, Vieira ML, Morais ZM, Vasconcellos SA, Nascimento AL: **A newly identified leptospiral adhesin mediates attachment to laminin.** *Infect Immun* 2006, **74**(11):6356-6364.
57. Verma A, Hellwage J, Artiushin S, Zipfel PF, Kraiczy P, Timoney JF, Stevenson B: **LfhA, a novel factor H-binding protein of Leptospira interrogans.** *Infect Immun* 2006, **74**(5):2659-2666.
58. Alexander RP, Zhulin IB: **Evolutionary genomics reveals conserved structural determinants of signaling and adaptation in microbial chemoreceptors.** *Proc Natl Acad Sci U S A* 2007, **104**(8):2885-2890.
59. Amin DN, Taylor BL, Johnson MS: **Organization of the aerotaxis receptor aer in the membrane of Escherichia coli.** *J Bacteriol* 2007, **189**(20):7206-7212.
60. Scheffers DJ, Pinho MG: **Bacterial cell wall synthesis: new insights from localization studies.** *Microbiol Mol Biol Rev* 2005, **69**(4):585-607.

Publish with **BioMed Central** and every scientist can read your work free of charge

*"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."*

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

