Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

HP-Lattice QSAR for dynein proteins: Experimental proteomics (2D-electrophoresis, mass spectrometry) and theoretic study of a *Leishmania infantum* sequence

María Auxiliadora Dea-Ayuela [a,b], Yunierkis Pérez-Castillo [c], Alfredo Meneses-Marcel [a,c], Florencio M. Ubeira [d], Francisco Bolas-Fernández [a], Kuo-Chen Chou [e], Humberto González-Díaz [d,e,*]

[a] *Department of Parasitology, Faculty of Pharmacy, Complutense University, 28040 Madrid, Spain*
[b] *Departamento de Atención Sanitaria, Salud Pública y Sanidad Animal, Facultad CC Experimentales y de La Salud, Universidad CEU Cardenal Herrera, 46113 Moncada (Valencia), Spain*
[c] *CBQ, Central University of Las Villas, 54830 Santa Clara, Cuba*
[d] *Department of Microbiology and Parasitology, Faculty of Pharmacy, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain*
[e] *Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA*

ARTICLE INFO

Article history:
Received 25 January 2008
Revised 23 June 2008
Accepted 2 July 2008
Available online 13 July 2008

Keywords:
QSAR
Spectral moments
HP-Lattice model
Complex networks
Sequence alignment
BLAST
Mass spectrometry
2D-Electrophoresis
Leishmania infantum
Parasites proteomics
Dyneins

ABSTRACT

The toxicity and inefficacy of actual organic drugs against Leishmaniosis justify research projects to find new molecular targets in *Leishmania* species including *Leishmania infantum* (*L. infantum*) and *Leishmania major* (*L. major*), both important pathogens. In this sense, quantitative structure–activity relationship (QSAR) methods, which are very useful in Bioorganic and Medicinal Chemistry to discover small-sized drugs, may help to identify not only new drugs but also new drug targets, if we apply them to proteins. Dyneins are important proteins of these parasites governing fundamental processes such as cilia and flagella motion, nuclear migration, organization of the mitotic splinde, and chromosome separation during mitosis. However, despite the interest for them as potential drug targets, so far there has been no report whatsoever on dyneins with QSAR techniques. To the best of our knowledge, we report here the first QSAR for dynein proteins. We used as input the Spectral Moments of a Markov matrix associated to the HP-Lattice Network of the protein sequence. The data contain 411 protein sequences of different species selected by ClustalX to develop a QSAR that correctly discriminates on average between 92.75% and 92.51% of dyneins and other proteins in four different train and cross-validation datasets. We also report a combined experimental and theoretic study of a new dynein sequence in order to illustrate the utility of the model to search for potential drug targets with a practical example. First, we carried out a 2D-electrophoresis analysis of *L. infantum* biological samples. Next, we excised from 2D-E gels one spot of interest belonging to an unknown protein or protein fragment in the region M < 20,200 and p*I* < 4. We used MASCOT search engine to find proteins in the *L. major* data base with the highest similarity score to the MS of the protein isolated from *L. infantum*. We used the QSAR model to predict the new sequence as dynein with probability of 99.99% without relying upon alignment. In order to confirm the previous function annotation we predicted the sequences as dynein with BLAST and the omniBLAST tools (96% alignment similarity to dyneins of other species). Using this combined strategy, we have successfully identified *L. infantum* protein containing dynein heavy chain, and illustrated the potential use of the QSAR model as a complement to alignment tools.

1. Introduction

Leishmania spp. are obligated intracellular protozoa that exist in two forms, a promastigote form (elongated cells with a long flagellum) and amastigote form (ovoid cells that have a very short flagellum). The flagellum is responsible for the motility of trypanosomatids and for their early interaction with the hosts, either by adhering to the insect digestive tract,[1] or by initiating the contact with mammalian cells.[2] Trypanosomatids depend on this adhesion to survive and differentiate.[3] This surface organelle plays a key role in *Leishmania* motility and sensory reception, and it is essential for parasite migration, invasion, and persistence host tissues.[4] The assembly and maintenance of flagellum require the continuous import of axoneme precursor proteins from the cytosol, as well as the removal of proteins generated by turnover

* Corresponding author. Tel.: +34 981 563100; fax: +34 981 594912.
E-mail address: gonzalezdiazh@yahoo.es (H. González-Díaz).

0968-0896/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2008.07.023

of axonemal structures. The import and export of these proteins appear to be largely mediated by intraflagellar transport particles that move along the axonemal doublet microtubules just beneath the flagellar membrane,[5] and are associated with either kinesin or dynein motor proteins, recycling kinesin, and discarding axoneme proteins back to the cytosol.[6] Dyneins contain one to three heavy chains, each one consisting of a C-terminal globular head and two elongated flexible structures called the stalk (the microtubule-binding domain) and the N-terminal tail (the cargo-binding domain). Dyneins are classified into two major categories, axonemal and cytoplasmatic.[7] The axonemal dynein produces the bending motion that propagates across cilia and flagella,[8] while cytoplasmatic dynein drives fundamental processes including nuclear migration, organization of the mitotic splinde, chromosome separation during mitosis, and the positioning and function of many intracellular organelles.[9] The dynein heavy chain, with a mass superior to 500 kDa, contains a 380 kDa motor domain in the C-terminal fragment,[10] incorporating sites for ATP hydrolysis and microtubule binding. Several studies have suggested that dynein is a member of the AAA$^+$ (ATPase associated with various cellular activities) superfamily.[11] Observations on dynein show six sequential linked AAA$^+$ ATPase-like domains, 35–40 kDa each,[12] but only the first domain binds and hydrolyzes ATP, while second domain is capable to bind ADP, possibly in a regulatory fashion.[13]

All these facts justify the search of new theoretical methods to facilitate the experimental discovery of dyneins, including new Computational Chemistry or Bioinformatics-inspired tools. In general, the great amount of information generated during the genomic and proteomic research is actually changing the way we think about molecular targets for the treatment of diseases. Consequently, many research laboratories are working in the direction of finding new molecular targets that remain unexploited up-to-date.[14] Various bioinformatics tools have been developed in order to successfully identify proteins using proteomic data, tools that have been reviewed by many authors.[15–18] On the other hand, quantitative structure–activity relationship (QSAR)[19–23] methodology has been successfully applied to small-sized molecules[24–29] and bio-macromolecules,[30–32] and specifically to proteins in Bioorganic and Medicinal Chemistry.[33–39] The QSAR methodology is based on the model search of connecting the protein function to numerical index characteristics of the protein structure. Some of the most useful indices used are known as topological indices (TIs) or connectivity indices (CIs).[40] TIs and/or CIs describe the connection by means of edges between nodes in complex networks (CNs) used to represent proteins. These CNs or graphs may represent a peptide, a protein, protein–protein relationships, whole proteomes, or even large macroscopic systems.[41–44] According to this picture, the nodes may be amino acids, 3D protein domains, or proteins, and the edges connecting nodes may be chemical bonds, structural similarity, functional relationships, or interactions between these nodes.[45] A very useful network model to represent one protein sequence is the so-called hydrophobicity-polarity lattice (HP-Lattice) model.[46,47] This raw but fast and useful representation has different variants, but essentially classifies all amino acids of one sequence as Hydrophobic and/or Polar and folds the sequence in a 2D or 3D Cartesian space, whose axes are the HP properties of amino acids and the sequence folds in one or the other direction according to the amino acid sequence assigned fixed length variation in coordinates of amino acids to form a compact Lattice network.[48–52] Next, one can use Statistical or Machine Learning techniques to process the CIs or TIs of these HP-Lattice networks and to find a relationship between the protein sequences and the biological function[53,54] in the same way as in classic QSAR for small-sized molecules.[55]

In this work, we introduced the first alignment-free QSAR model capable of identifying whether a protein contains dynein heavy chains. This model was challenged to correctly classify a putative dynein fragment from a 2D-electrophoresis (2D-E) spot of *Leishmania infantum* (*L. infantum*) identified by mass spectrometry (MS). The work opens a door for the integrated use of experimental techniques and Bioinformatics with Computational Chemistry on the study of isolated proteins as *Leishmania* and other important pathogens.

2. Results and discussion

2.1. HP-Lattice network-based QSAR study

In order to complement BLAST results, we can develop an alignment-free model depending on sequence descriptors to recognize proteins with a given activity.[56] Other authors have used sequence information combined with structural information to predict new dynein-related proteins.[57] However, to the best of our knowledge, there has not been reported any QSAR study of this family. In this work, we introduce for the first time a QSAR model able to differentiate proteins containing dynein heavy chain domain from those that do not contain it, based on the differences between the HP-Lattices of the dyneins and no-dyneins sequences. In Figure 1 we depict the superposition of all HP-Lattice networks for dyneins (grey color) and no-dynein proteins (black). We can detect some visual differences between the two groups of proteins (dyneins tend to fold in the center and no-dyneins in the exterior of the 2D Cartesian space). However, a quantitative analysis is necessary for a final assignment of a protein.

Consequently, we calculated the π_k values for all the sequences and carried out a LDA. LDA has been selected not only due to its simplicity but also to its many applications in QSAR[58] and proteomics.[59] The QSAR discriminant equation obtained has four parameters.

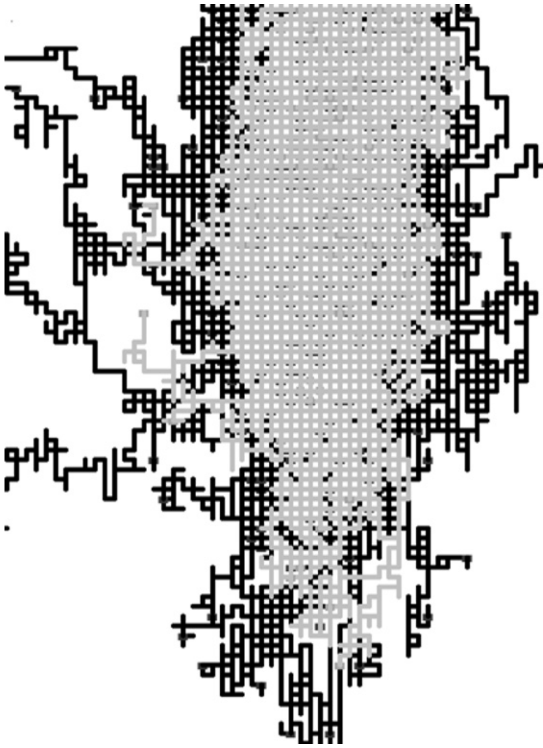


Figure 1. Superposition of HP-Lattice networks for dyneins (gray color) and no-dynein proteins (black).

$$\text{dynein-score} = -120.59 \cdot \pi_s + 43.02 \cdot \pi_o - 159.71 \cdot \pi_1 + 236.36 \cdot \pi_3 + 0.01$$
$$N = 411, \quad U = 0.36, \quad F = 181.80, \quad p\text{-level} < 0.001, \qquad (1)$$

where N is the number of proteins used to seek the model, U is Wilk's lambda or U-statistic, F is Fischer ratio, and p is the p-level.[60–62] The model (1) correctly classifies 192 of 211 (91.00%) dyneins and 200 of 210 (95.24%) no-dyneins for training series. For the cross-validation or prediction series, 128 of 140 (91.43%) are correctly classified. More specifically, 62/70 dyneins and 66/70 no-dyneins are well classified. The present results are considered very good for this type of LDA-based QSAR classification models reported by different researchers in QSAR.[63–65] Additionally, a re-substitution procedure was carried out by interchanging proteins in training and prediction series.[33] Results for re-substitution procedure are shown in *Supplementary* Information (see Tables SI1–SI3). From *Table* SI1, the high robustness and predictability of the model can be noted. As a result of the re-substitution procedure, in average, 92.75% of proteins in training series and 92.50% of those in validation series are correctly classified. These results clearly show that our model is capable of differentiating proteins which contain dynein heavy chains in their sequence from those that do not, with a high-level of accuracy.

2.2. Experimental case study

In this section we presented an example of using the above QSAR model to predict the presence of dynein heavy chain in a protein sequence obtained from *L. infantum*. In Figure 2, we illustrate an overall view of the 2D-E map obtained from the *L. infantum* promastigote homogenate. In this figure we have done a zooming in the left-to-down corner to highlight an area of high density of spots, which apparently corresponds to protein fragments of low MW and low p*I*. Our interest in this area derived from the fact that these spots remained invariable from gel to gel repetitions and might correspond to relevant proteins of this parasite. Initially, we selected three spots marked with numbers from 1 to 3 within small white circles as shown in the zooming area of Figure 2, in order to start the investigation of the nature of these proteins. The protein contained in each spot was submitted to in-gel trypsin digestion, and the mass of the resulting peptides was obtained from MALDI-TOF MS analysis. However, since in subsequent studies we have observed that only the protein contained in spot #2 corresponded to a dynein, we will only refer to the steps we follow for the identification of this *L. infantum* protein as the objective of this study.

Once we obtained the data from MALDI-TOF MS analysis of spot #2, the most relevant MS signals were introduced into the MASCOT search engine.[66,67] Due to the fact that the MASCOT collection of annotated databases did not contain data about *L. infantum* proteome, we chose the *Leishmania major* database of annotated proteins with MS recorded because of its similarity to *L. infantum*.[68] Even if a protein fragment has a low MW, the MASCOT search of MS signals found two hits with a score higher than 51 ($p < 0.05$) for spot #2 (see Table 1 and Fig. 3). The top score found was 57, correspondent to the protein CHR26_tmp.127 of *L. major* with 460,680 Da mass and described as dhc16f protein. This protein is assigned as possible *L. major* protein but without function annotation. This protein sequence in FASTA[69] format is reported in *Supplementary* Information (see Note SI1).

Table 1
Top-five hits found for the unknown electrophoresis dot with the MASCOT analysis

Hit number	Protein	Mass	Score[a]	Protein description
1	CHR26_tmp.127	460 680	57	Unknown function
2	CHR33_tmp.135	169 257	56	Unknown function
3	LmjF31.2850c	22 350	39	Ribosomal protein
4	CHR30_tmp.180c	24 128	39	Ribosomal protein
5	LmjF36.4550	129 291	39	Elongation factor-2 kinase efk-1b isoform

[a] Significant proteins are those with scores higher than 51.

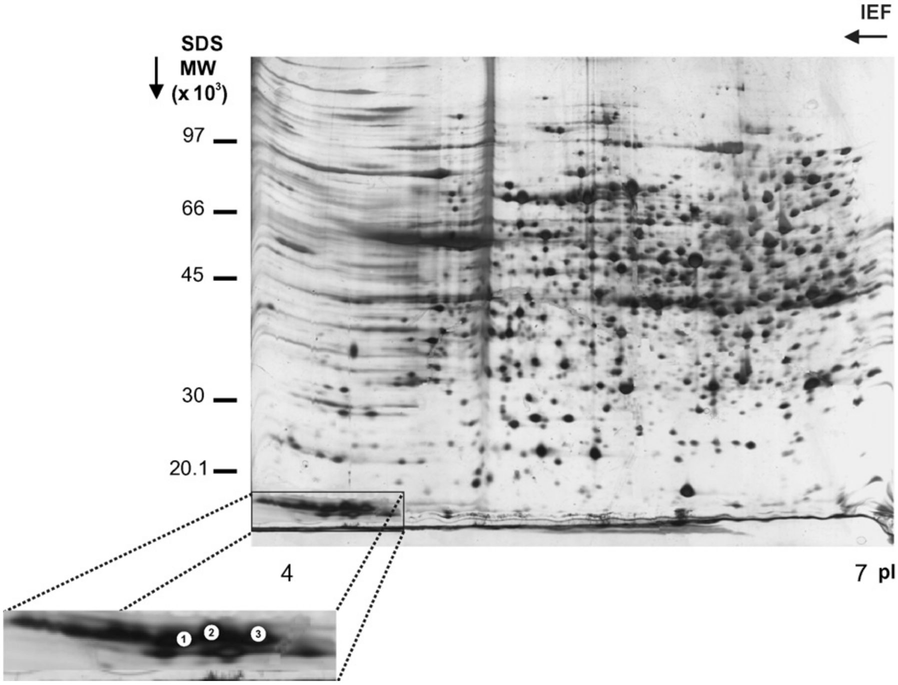


Figure 2. 2-DE analysis of proteins from *L. infantum* promastigotes. IEF was performed with 800 µg of proteins using a 4–7 pH range strips. SDS–PAGE was performed on 12.5% polyacrylamide gels and stained with silver. On the left a zooming of the area with spots of interest.

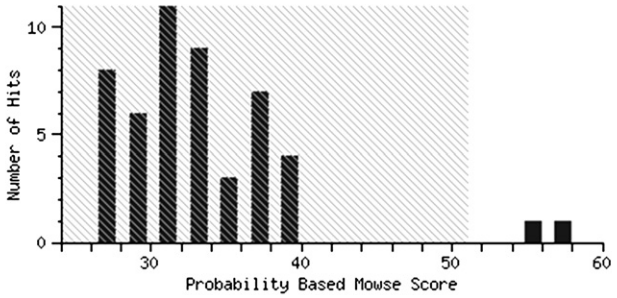


Figure 3. Distribution of hits found using the MASCOT search engine.

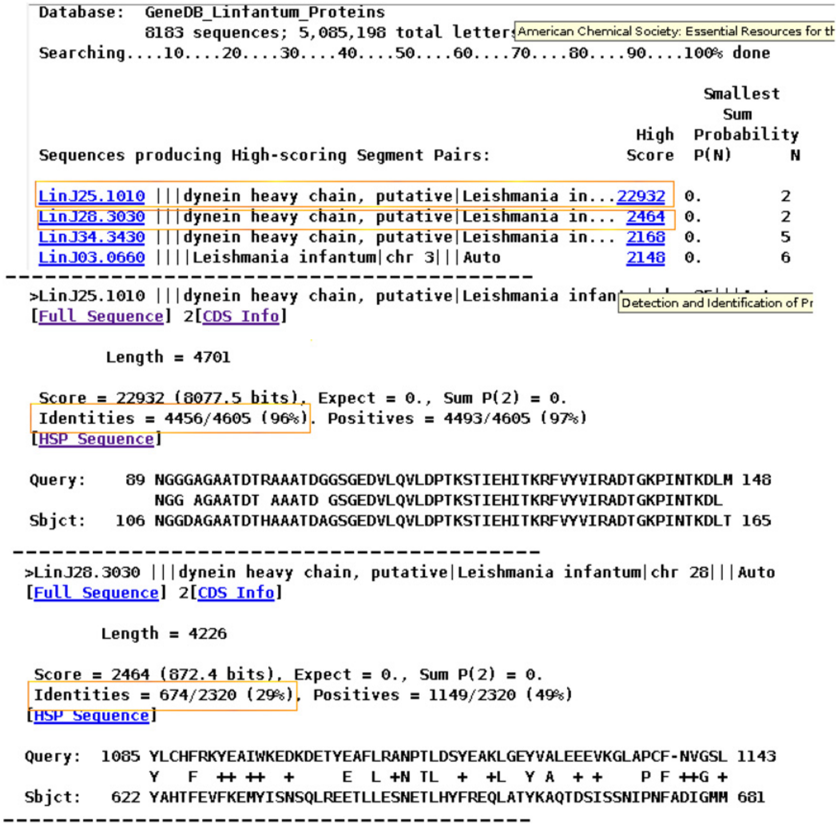
In *Supplementary* Material (Table SI4) we provide the summary results for sequence assignment of the most important matching MS signals using the MASCOT approach for this protein and the sample. The largest MS signal identified was 2163.06 Da that is coincident with experimental signal 2163.08, which corresponded to the peptide sequence SIEDKLATLQADMEENVLK. There was no match in the MASCOT analysis to signals: 781.53, 789.39, 823.45, 839.43, 855.03, 855.45, 861.04, 877.46, 971.50, 1019.51, 1364.69, 1437.68, 1449.86, 1454.73, 1537.91, 1666.87, 1688.84, 1758.01, 2185.05, and 2201.01 Da. The protein CHR33_tmp.135 with mass 169,257 and score 56 was the second and last significant hit found (see also Fig. 3). This protein is also recorded as possible *L. major* protein but also appears without a known function annotated. In *Supplementary* Material (Table SI4) we also provide the summary results for sequence assignment of the most important matching MS signals using the MASCOT approach for this second protein and the sample. No matching MS signals for this second protein and the samples are 773.48, 779.47, 781.53, 823.45, 839.43, 839.43, 855.03, 855.45, 861.04, 867.48, 877.46, 883.47, 905.52, 965.55, 971.50, 1002.53, 1364.69, 1449.86, 1454.73, 1455.75, 1483.71, 1493.88, 1581.92, 1666.87, 1669.97, 1714.00, 1758.01, 2163.06, 2185.05, and 2201.01 Da. In Supplementary Information file we give details for all peptides sequences corresponding to these MS signals.

2.3. BLAST analysis

We carried out a BLAST search, using the omniBLAST tool available at the Sanger Institute, using as query database the *L. infantum* genome database, in order to obtain more information about the function of our protein. The Sanger Institute makes public available annotated databases of the genome of different organisms.[70] A summary of omniBLAST results is shown in Figure 4. As it can be seen, according to omniBLAST results, we have found one protein in the *L. infantum* database which is very similar (96% of identity) to *L. major* previously identified using the MASCOT server. This result could indicate that the protein previously isolated using 2D-E (spot #2) may correspond to the entry LinJ25.1010 in the database of *L. infantum* genome. As expected, when transferring annotations from an organism to another, the protein LinJ25.1010 is also annotated to have/present a dynein function.

2.4. QSAR prediction of new dyneins

After provisional assignment of the protein fragment (spot #2) as a dynein, we decided to use the above reported model for dyneins to predict whether the entries found with MASCOT and their homologues gi|68126919, which correspond to the *L. major* protein and LinJ25.1010 corresponding to *L. infantum*, contain a dynein heavy chain. Calculating the CIs (π_k values) for these sequences in-


```
 Database:   GeneDB_Linfantum_Proteins
             8183 sequences; 5,085,198 total letter American Chemical Society: Essential Resources for th
 Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                  Smallest
                                                                    Sum
                                                            High  Probability
 Sequences producing High-scoring Segment Pairs:            Score  P(N)      N

 LinJ25.1010 |||dynein heavy chain, putative|Leishmania in...22932  0.        2
 LinJ28.3030 |||dynein heavy chain, putative|Leishmania in... 2464  0.        2
 LinJ34.3430 |||dynein heavy chain, putative|Leishmania in... 2168  0.        5
 LinJ03.0660 ||||Leishmania infantum|chr 3|||Auto             2148  0.        6
----------------------------------------
 >LinJ25.1010 |||dynein heavy chain, putative|Leishmania infan Detection and Identification of Pr
 [Full Sequence] 2[CDS Info]

        Length = 4701

  Score = 22932 (8077.5 bits), Expect = 0., Sum P(2) = 0.
  Identities = 4456/4605 (96%), Positives = 4493/4605 (97%)
 [HSP Sequence]

 Query:    89 NGGGAGAATDTRAAATDGGSGEDVLQVLDPTKSTIEHITKRFVYVIRADTGKPINTKDLM 148
              NGG AGAATDT AAATD GSGEDVLQVLDPTKSTIEHITKRFVYVIRADTGKPINTKDL
 Sbjct:   106 NGGDAGAATDTHAAATDAGSGEDVLQVLDPTKSTIEHITKRFVYVIRADTGKPINTKDLT 165

 ----------------------------------------
 >LinJ28.3030 |||dynein heavy chain, putative|Leishmania infantum|chr 28|||Auto
 [Full Sequence] 2[CDS Info]

        Length = 4226

  Score = 2464 (872.4 bits), Expect = 0., Sum P(2) = 0.
  Identities = 674/2320 (29%), Positives = 1149/2320 (49%)
 [HSP Sequence]

 Query:  1085 YLCHFRKYEAIWKEDKDETYEAFLRANPTLDSYEAKLGEYVALEEEVKGLAPCF-NVGSL 1143
              Y   F  ++ ++  +      E  L +N TL  +  +L  Y A  + +    P F ++G +
 Sbjct:   622 YAHTFEVFKEMYISNSQLREETLLESNETLHYFREQLATYKAQTDSISSNIPNFADIGMM 681
----------------------------------------
```

Figure 4. Fragment of the omniBLAST result using as query sequence that of the entry gi|68126919. Note the difference in similarity of the best two hits (marked as orange rectangles).

cluded in the discriminant equation and using the obtained QSAR model, we predicted gi|68126919 with a 99.99% probability to belong to the group of dynein proteins. The same result was obtained evaluating the sequence of the entry LinJ25.101. These results are very significant for a case predicted with this kind of LDA-based QSAR models, based on TIs or CIs[71,72].

3. Conclusions

In this work, we present the first alignment-free QSAR model to determine whether a protein presents a dynein heavy chain, using only the protein sequence as input data. In this study, we have shown that the CIs derived from HP-Lattice networks encode important biological information to predict dyneins without relying upon sequence alignment. In order to exemplify the use of the QSAR model, we have also used the combination of Proteomic (2D-E and MS), Computational Chemistry (QSAR), and Bioinformatics (BLAST, MASCOT, omniBLAST) techniques with the aim to experimentally identify and predict a protein from *L. infantum.* The work follows the same line of some of our previous results[73–75] attracting researchers' attention on the potential uses of the protein study of extended QSAR, as a complement to alignment techniques.

4. Materials and methods

4.1. Computational methods

4.1.1. QSAR analysis

First, the corresponding protein sequences were compelled to fold in a 2D Cartesian space to form a HP-Lattice network. Next, we calculated the alignment-free CIs for each protein. We used a Markov model (MM) to calculate the CIs of each protein and codify information about protein sequences.[76] Specifically, we used the CIs called Spectral Moments of the Markov matrix[77] (π_k), in this work. The procedure has been published in detail elsewhere,[78,79] so we omit herein a detailed mathematical description. The aim of this QSAR study was to construct a linear discriminant analysis (LDA)[80] model able to classify proteins into two groups, based on their sequences represented by the π_k values. We defined a 'positive group' as the proteins that have a dynein function previously demonstrated and a 'negative group' as the proteins that have functions other than dynein activity. We carried out LDA and all statistical analyses using the software STATISTICA.[81] With this software we obtained a discriminant function QSAR model in the following form.

$$\text{dynein-score} = a_k \cdot \pi_k + \cdots + a_2 \cdot \pi_2 + a_1 \cdot \pi_1 + a_0 \cdot \pi_0 + a, \qquad (2)$$

where dynein-score is introduced as a dummy variable to fit the model with values dynein-score = 1 for dynein proteins and dynein-score = -1 for 'negative group' of no-dynein proteins. In order to choose the dynein proteins group for QSAR study, we have used the CDART available at NCBI.[82] For this selection we used as query sequence the protein previously identified from the analysis of MS data by MASCOT search engine. On the other hand, the 'negative' group of no-dynein proteins is extracted from the ExProt database,[83] which is a database containing proteins with experimentally determined functions. The negative group is selected so that it contains as high functional diversity as possible, having the same amount of proteins as the 'positive' one and where every protein has a number of amino acids in the same range as in the other group.

4.1.2. QSAR input data selection

Querying the protein sequence found with MASCOT against the Conserved Domain Architecture Retrieval Tool (CDART)[82] retrieved 328 proteins containing dynein heavy chain. Since our main purpose is to predict the presence of dynein heavy chain, and for the sake of calculation simplicity, we have aligned all sequences using the ClustalX[84] software with the aim to select, for every protein, only the region which corresponds to the dynein heavy chain domain. We chose, as the representative sequence for region selection, the entry XP_804816 corresponding to the *Trypanosoma cruzi* dynein heavy chain, based on the similarity between parts of the proteome of the three trypanosomatids *L. major, Trypanosoma brucei,* and *T. cruzi.*[85] Once the regions are selected, to avoid including short or extremely long segments of proteins, we selected only those sequence segments that are between 500 and 800 amino acids long. As a consequence of this selection process, 281 sequence segments are finally included in the 'active' group. For the 'inactive' group conformation, we randomly extracted 280 proteins from the Exprot database; the only criteria used during selection was not to have less than 500 and more than 800 amino acids, and proteins included in this group had a functional diversity as high as possible. Codes for proteins which segments are included in the dynein ('active') and not dynein ('inactive') group are listed in Supporting Information. Functions annotated in the ExProt database for proteins included in the 'inactive' group and sequences for proteins fragments used during model construction are also included in Supporting Information.

4.1.3. Database search

The peptide mass fingerprinting data, obtained from MALDI-TOF MS analyses, were used to search for protein candidates in two sequence databases: SWISS-PROT/TrEMBL non-redundant protein database (www.expasy.ch/sprot) and a complete genomic database from the related species *Leishmania major*, namely, ftp://ftp.sanger.ac.uk/pub/databases/L.major_sequences/LEISHPEP/, using MASCOT software program (www.matrixscience.com). The MASCOT search parameters were adjusted according to the MS experiment carried out and to the above description as follows: type of search: Sequence Query; enzyme: trypsin; fixed modifications: carbamidomethyl (C); variable modifications: oxidation (M); mass values: monoisotopic; protein mass: unrestricted; peptide mass tolerance: ±100 ppm; fragment mass tolerance: ±0.4 Da; max missed cleavages: 1; instrument type: MALDI-TOF-TOF. We introduced the MS signals corresponding to one of the unidentified 2D-electrophoresis spots (protein) into the MASCOT analysis system. The sample was recorded on this web page with the search title: Sample Set ID: 1122, AnalysisID: 1466, Maldi WellID: 17500, Spectrum ID: 7971, Path=\040519*Leishmania*\New Analysis 2. The database used was *Leishmania* 290 703 (with 7467 sequences; and 4,469,604 residues).

4.1.4. BLAST search of putative protein function

For the prediction of this protein function, we introduced the sequence of CHR26_tmp.127 into the BLAST analysis to get an initial idea of its function.[86] The BLAST procedure was carried out using as query database the non-redundant NCI database and allowing BLAST to search for conserved domains through the CD search tool.[87]

4.2. Experimental methods

4.2.1. Cell culture of parasites

Promastigotes of the *Leishmania* strain LEM75 were grown in medium Schneider supplemented to a final concentration of 0.4 g/l $NaHCO_3$, 4 g/l Hepes,100 mg/l penicillin and 100 mg/l streptomycin, and 10% fetal bovine serum (Gibco), pH 6.8, and 26 °C.

4.2.2. Sample preparation

Mid-log promastigotes were recovered on the seventh day post-inoculum (pi) and the parasites were centrifuged at 3000 rpm for

10 min at 4 °C. The resulting pellet was washed five times with Tris–HCl, pH 7.8, and resuspended in 0.1 ml of this same buffer. The sample was sonicated for 10 s with a Virsonic 5 (Virtis, NY, USA) set at 70% output power in ice bath. The homogenate was extracted in 5 mM Tris–HCl buffer, pH 7.8, containing 1 mM phenylmethylsulfonyl fluoride (PMSF) as a protease inhibitor, at 4 °C overnight and, subsequently centrifuged at 10,000*g* for 1 h at 4 °C (Biofuge 17RS: Heraeus Sepatech, Gmb, Osterode, Denmark). The supernatant was dialysed overnight at 4 °C in 0.5 mM Tris–HCl buffer. Proteins were precipitated with 20% TCA (trichloroacetic acid) in acetone with 20 mM DTT for 1 h at −20 °C, and were added at a ratio of 1:1 to the homogenated. Then, the sample was centrifuged at 10,000*g* for 15 min and the pellet was washed with cold acetone containing 20 mM DTT. Residual acetone was removed by air-drying. In order to achieve a well-focused first-dimension separation, sample proteins must be completely disaggregated and fully solubilized, in a sample buffer containing 7 M urea, 2 M thiourea, 4% Chaps, Destreak buffer (Amersham Bioscience), 5 mM CO_3K_2, 2% IPG buffer (Amersham Bioscience), and incubated at room temperature for 30 min. Following clarification by centrifugation at room temperature (12,000*g*, 10 min), the supernatant was stored frozen at −20 °C.

4.2.3. 2D-E experiments

In total 340 μl of rehydration buffer was added to promastigotes solubilized extracts (7 M urea, 2 M thiourea, 2% Chaps, 0.75% IPG buffer 4–7, bromophenol blue) and were immediately adsorbed onto 18 cm immobilized pH 4–7 gradient (IPG) strips (Amersham Biosciences).[88] Optimal IEF was carried out at 20 °C, with an active rehydration step of 12 h (50 V), and then focused on an IPGphor IEF unit (Amersham Biosciences) by using the following program: 150 V for 2 h, 500 V for 1 h, 1000 V for 1 h, 1000–2000 V for 1 h and 8000 V for 6 h. After focusing, IPG strips were equilibrated for 15 min in 10 ml of 50 mM Tris–HCl, pH 8.8, 6 M urea, 30% v/v glycerol, 2% w/v SDS, traces of bromophenol blue, and 100 mg of DTT. Then, the strips were incubated for 25 min in the same buffer but replacing DTT by 300 mg of iodoacetamide. After equilibration, the IPG strips were placed onto 12.5% SDS–polyacrylamide gels and sealed with 0.5% (w/v) agarose. SDS–PAGE was run at 15 mA/gel. 2D gels were stained with silver staining mass spectrometry compatible. Briefly, the gels were fixed in 40% ethanol (v/v), 10% (v/v) acetic acid overnight, then were sensitized with sodium acetate 0.68% (w/v) and 0.05% sodium thiosulfate for 30 min and washed with deionized water three times for 5 min. The gels were incubated in 0.25% (w/v) silver nitrate for 30 min. After incubation, it was rinsed with deionized water two times for 50 s followed by adding the developing solution, which contained 2.5% (w/v) sodium carbonate with 0.04% (v/v) formaldehyde until the desired intensity range. Development was finished by adding 1.5% (w/v) EDTA.

4.2.4. MALDI-TOF MS experiment

Spots of interest were manually excised from silver stained 2-DE gels after being destained, as described by Gharahdaghi et al.[89] Then, gel pieces were incubated with 12.5 ng/μl sequencing grade trypsin (Roche Molecular Biochemicals) in 25 mM AmBic, overnight, at 37 °C. After digestion, the supernatants (crude extracts) were separated. Peptides were extracted from the gel pieces first into 50% ACN, 1% trifluoroacetic acid and then into 100% ACN. Then, one microliter of each sample and 0.4 μl of 3 mg/ml α-cyano-4-hydroxycinnamic acid matrix (Sigma) in 50% ACN, and 0.01% trifluoroacetic acid were spotted onto a MALDI target. MALDI-TOF MS analyses were performed on a Voyager-DE STR mass spectrometer (PerSeptive Biosystems, Framingham, MA, USA). The following parameters were used: cysteine as *s*-carbamidomethyl derivative and methionine in oxidized form. Spectra were acquired over the *m/z* range of 700–4500 Da.

Acknowledgments

First of all, the authors thank the Editor Herbert Waldmann and two unknown referees for their kind attention and comments. This work was partially funded by Grant No. 08.3/0019.1/2001 of the Comunidad Autnónoma de Madrid. M.A.D.A. was a Post-doctoral Fellow of the Comunidad Autónoma de Madrid. González-Díaz, H. acknowledges a tenure-track research contract at the Faculty of Pharmacy, USC, funded by Program Isidro Parga Pondal of the Xunta de Galicia. González-Díaz, H. thanks Xunta de Galicia and European Union (F.S.E.) Grant IN809A 2007/84-0 for one-year post-doctoral training supervised by Prof. Dr. Kuo-Chen Chou at the Gordon Life Science Institute, CA, USA.

Supplementary data

Supporting Information associated to this paper is published on line, on the journal web site and contains: the accession codes for proteins, function, predicted probability, and training or validation group for all proteins used in the QSAR study. Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2008.07.023.

References

1. Vickerman, K.; Tetley, L. In *Ciliary and Flagellar Membranes*; Bloodgood, R. A., Ed.; Booknews: London, 1990; p 267.
2. De Souza, W. *Int. Rev. Cytol.* **1984**, *86*, 197.
3. De Souza, W. *Prog. Protistol.* **1989**, *3*, 87.
4. Landfear, S. M.; Ignatushchenko, M. *Mol. Biochem. Parasitol.* **2001**, *115*, 1.
5. Rosenbaum, J. L.; Witman, G. B. *Nat. Rev. Mol. Cell Biol.* **2002**, *3*, 813.
6. Tull, D.; Vince, J. E.; Callaghan, J. M.; Naderer, T.; Spurck, T.; McFadden, G. I.; Currie, G.; Ferguson, K.; Bacic, A.; McConville, M. J.*Mol. Biol. Cell* **2004**, *15*, 4775.
7. Oiwa, K.; Sakakibara, H. *Curr. Opin. Cell Biol.* **2005**, *17*, 98.
8. Di Bella, L. M.; King, S. M. *Int. Rev. Cytol.* **2001**, *210*, 227.
9. Vale, R. D. *Cell* **2003**, *112*, 467.
10. Nishiura, M.; Kon, T.; Shiroguchi, K.; Ohkura, R.; Shima, T.; Toyoshima, Y. Y.; Sutoh, K. *J. Biol. Chem.* **2004**, *279*, 22799.
11. Ogura, T.; Wilkinson, A. J. *Genes Cells* **2001**, *6*, 575.
12. Burgess, S. A.; Walker, M. L.; Sakakibara, H.; Knight, P. J.; Oiwa, K. *Nature* **2003**, *421*, 715.
13. Yagi, T. *Cell Struct. Funct.* **2000**, *25*, 263.
14. Drews, J. *Drug Discovery Today* **2003**, *8*, 411.
15. Huynen, M. A.; Snel, B.; von Mering; Brok, P. *Curr. Opin. Cell Biol.* **2003**, *15*, 191.
16. Chou, K. C. *Curr. Med. Chem.* **2004**, *11*, 2105.
17. Sirois, S.; Wei, D. Q.; Du, Q.; Chou,K. C. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1111.
18. Chou, K. C.; Wei, D. Q.; Du, Q. S.; Sirois, S.; Zhong, W. Z. *Curr. Med. Chem.* **2006**, *13*, 3263.
19. Du, Q. S.; Huang, R. B.; Wei, Y. T.; Wang, C. H.; Chou, K. C. *J. Comput. Chem.* **2007**, *28*, 2043.
20. Du, Q. S.; Huang, R. B.; Wei, Y. T.; Du, L. Q.; Chou, K. C. *J. Comput. Chem.* **2008**, *29*, 211.
21. Rasulev, B. F.; Saidkhodzhaev, A. I.; Nazrullaev, S. S.; Akhmedkhodzhaeva, K. S.; Khushbaktova, Z. A.; Leszczynski, J. *SAR QSAR Environ. Res.* **2007**, *18*, 663.
22. Isayev, O.; Rasulev, B.; Gorb, L.; Leszczynski, J. *Mol. Divers.* **2006**, *10*, 233.
23. Dyguda, E.; Grembecka, J.; Sokalski, W. A.; Leszczynski, J. *J. Am. Chem. Soc.* **2005**, *127*, 1658.
24. Dixon, S.; Merz, K. M., Jr.; Lauri, G.; Ianni, J. C. *J. Comput. Chem.* **2005**, *26*, 23.
25. Du, Q.; Mezey, P. G.; Chou, K. C. *J. Comput. Chem.* **2005**, *26*, 461.
26. Castillo-Garit, J. A.; Marrero-Ponce, Y.; Torrens, F. *Bioorg. Med. Chem.* **2006**, *14*, 2398.
27. Castillo-Garit, J. A.; Marrero-Ponce, Y.; Torrens, F.; Rotondo, R. *J. Mol. Graphics Modell.* **2007**, *26*, 32.
28. Duchowicz, P. R.; Fernandez, M.; Caballero, J.; Castro, E. A.; Fernandez, F. M. *Bioorg. Med. Chem.* **2006**, *14*, 5876.
29. Caballero, J.; Garriga, M.; Fernandez, M. *Bioorg. Med. Chem.* **2006**, *14*, 3330.
30. Hua, S.; Sun, Z. *Bioinformatics* **2001**, *17*, 721.
31. Randic, M.; Balaban, A. T. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 532.
32. Du, Q. S.; Wei, Y. T.; Pang, Z. W.; Chou, K. C.; Huang, R. B. *Protein Eng. Des. Sel.* **2007**, *20*, 417.
33. Gonzalez-Diaz, H.; Molina, R.; Uriarte, E. *FEBS Lett.* **2005**, *579*, 4297.
34. Ramos de Armas, R.; Gonzalez Diaz, H.; Molina, R.; Uriarte, E. *Proteins* **2004**, *56*, 715.
35. Zhou, H.; Zhou, Y. *Proteins* **2002**, *49*, 483.

36. Fernández, M.; Caballero, J.; Fernández, L.; Abreu, J. I.; Garriga, M. *J. Mol. Graphics Modell.* **2007**, *26*, 748.
37. Fernández, L.; Caballero, J.; Abreu, J. I.; Fernández, M. *Proteins* **2007**, *67*, 834.
38. Caballero, J.; Fernández, L.; Garriga, M.; Abreu, J. I.; Collina, S.; Fernández, M. *J. Mol. Graphics Modell.* **2007**, *26*, 166.
39. Caballero, J.; Fernandez, L.; Abreu, J. I.; Fernandez, M. *J. Chem. Inf. Model.* **2006**, *46*, 1255.
40. González-Díaz, H.; Vilar, S.; Santana, L.; Uriarte, E. *Curr. Top. Med. Chem.* **2007**, *7*, 1025.
41. Estrada, E. *Proteomics* **2006**, *6*, 35.
42. Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M.; Hwang, D. U. *Phys. Rep.* **2006**, *424*, 175.
43. Vazquez, A.; Dobrin, R.; Sergi, D.; Eckmann, J. P.; Oltvai, Z. N.; Barabasi, A. L. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 17940.
44. Albert, R.; Jeong, H.; Barabasi, A. L. *Nature* **2000**, *406*, 378.
45. González-Díaz, H.; González-Díaz, Y.; Santana, L.; Ubeira, F. M.; Uriarte, E. *Proteomics* **2008**, *8*, 750.
46. Agarwala, R.; Batzoglou, S.; Dancik, V.; Decatur, S. E.; Hannenhalli, S.; Farach, M.; Muthukrishnan, S.; Skiena, S. *J. Comput. Biol.* **1997**, *4*, 275.
47. Berger, B.; Leighton, T. *J. Comput. Biol.* **1998**, *5*, 27.
48. Yu, Z. G.; Anh, V.; Lau, K. S. *J. Theor. Biol.* **2004**, *226*, 341.
49. Thachuk, C.; Shmygelska, A.; Hoos, H. H. *BMC Bioinform.* **2007**, *8*, 342.
50. Gupta, A.; Manuch, J.; Stacho, L. *Proc. IEEE Comput. Syst. Bioinform. Conf.* **2004**, 311.
51. Gupta, A.; Manuch, J.; Stacho, L. *J. Comput. Biol.* **2005**, *12*, 1328.
52. Jiang, M.; Zhu, B. *J. Bioinform. Comput. Biol.* **2005**, *3*, 19.
53. Agüero-Chapin, G.; Gonzalez-Diaz, H.; Molina, R.; Varona-Santos, J.; Uriarte, E.; Gonzalez-Diaz, Y. *FEBS Lett.* **2006**, *580*, 723.
54. Agüero-Chapín, G.; González-Díaz, H.; de la Riva, G.; Rodríguez, E.; Sánchez-Rodríguez, A.; Podda, G.; Vazquez-Padrón, R. I. *J. Chem. Inf. Model.* **2008**, *48*, 434.
55. Vilar, S.; Estrada, E.; Uriarte, E.; Santana, L.; Gutierrez, Y. *J. Chem. Inf. Model.* **2005**, *45*, 502.
56. Han, L.; Cui, J.; Lin, H.; Ji, Z.; Cao, Z.; Li, Y.; Chen, Y. *Proteomics* **2006**, *6*, 4023.
57. Parisi, G.; Fornasari, M. S.; Echave, J. *FEBS Lett.* **2004**, *562*, 1.
58. Marrero-Ponce, Y.; Khan, M. T.; Casanola Martin, G. M.; Ather, A.; Sultankhodzhaev, M. N.; Torrens, F.; Rotondo, R. *ChemMedChem* **2007**, *2*, 449.
59. Lilien, R. H.; Farid, H.; Donald, B. R. *J. Comput. Biol.* **2003**, *10*, 925.
60. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Olazabal, E.; Serrano, H. S.; Morales, A.; Castanedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Sanchez, A. M.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2005**, *13*, 1005.
61. Garcia-Garcia, A.; Galvez, J.; de Julian-Ortiz, J. V.; Garcia-Domenech, R.; Munoz, C.; Guna, R.; Borras, R. *J. Antimicrob. Chemother.* **2004**, *53*, 65.
62. Patankar, S. J.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 885.
63. Casanola-Martin, G. M.; Marrero-Ponce, Y.; Khan, M. T.; Ather, A.; Sultan, S.; Torrens, F.; Rotondo, R. *Bioorg. Med. Chem.* **2007**, *15*, 1483.
64. Casanola-Martin, G. M.; Marrero-Ponce, Y.; Khan, M. T.; Ather, A.; Khan, K. M.; Torrens, F.; Rotondo, R. *Eur. J. Med. Chem.* **2007**, *42*, 1370.
65. Alvarez-Ginarte, Y. M.; Marrero-Ponce, Y.; Ruiz-Garcia, J. A.; Montero-Cabrera, L. A.; Vega, J. M.; Noheda Marin, P.; Crespo-Otero, R.; Zaragoza, F. T.; Garcia-Domenech, R. *J. Comput. Chem.* **2007**.
66. Lei, Z.; Elmer, A. M.; Watson, B. S.; Dixon, R. A.; Mendes, P. J.; Sumner, L. W.*Mol. Cell. Proteomics* **2005**, *4*, 1812.
67. Giddings, M. C.; Shah, A. A.; Gesteland, R.; Moore, B. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 20.
68. Arakaki, T.; Le Trong, I.; Phizicky, E.; Quartley, E.; DeTitta, G.; Luft, J.; Lauricella, A.; Anderson, L.; Kalyuzhniy, O.; Worthey, E.; Myler, P. J.; Kim, D.; Baker, D.; Hol, W. G.; Merritt, E. A. *Acta Crystallogr., Sect. F Struct. Biol. Cryst. Commun.* **2006**, *62*, 175.
69. Durand, P.; Canard, L.; Mornon, J. P. *Comput. Appl. Biosci.* **1997**, *13*, 407.
70. Parkinson, J.; Whitton, C.; Schmid, R.; Thomson, M.; Blaxter, M. *Nucleic Acids Res.* **2004**, *32*, D427.
71. Gozalbes, R.; Brun-Pascaud, M.; Garcia-Domenech, R.; Galvez, J.; Pierre-Marie, G.; Jean-Pierre, D.; Derouin, F. *Antimicrob. Agents Chemother.* **2000**, *44*, 2771.
72. Gozalbes, R.; Galvez, J.; Garcia-Domenech, R.; Derouin, F. *SAR QSAR Environ. Res.* **1999**, *10*, 47.
73. González-Díaz, H.; Saiz-Urra, L.; Molina, R.; Santana, L.; Uriarte, E. *J. Proteome Res.* **2007**, *6*, 904.
74. González-Díaz, H.; Saiz-Urra, L.; Molina, R.; Gonzalez-Diaz, Y.; Sanchez-Gonzalez, A. *J. Comput. Chem.* **2007**, *28*, 1042.
75. González-Díaz, H.; Pérez-Castillo, Y.; Podda, G.; Uriarte, E. *J. Comput. Chem.* **2007**, *28*, 1990.
76. Freund, J. A.; Poschel, T. In *Lecture Notes in Physics*; Springer: Berlin, Germany, 2000.
77. Gonzalez, M. P.; del Carmen Teran Moldes, M. *Bioorg. Med. Chem.* **2004**, *12*, 2985.
78. Agüero-Chapin, G.; Gonzalez-Dıaz, H.; Molina, R.; Varona-Santos, J.; Uriarteb, E.; Gonzalez-Dıaz, Y. *FEBS Lett.* **2006**, *580*, 723.
79. Gonzalez-Dıaz, H.; Ramos de Armas, R.; Molina, R. *Bioinformatics* **2003**, *19*, 2079.
80. Van Waterbeemd, H. In *Method and Principles in Medicinal Chemistry*, Manhnhold, R., Krogsgaard-Larsen, P., Timmerman, H., Van Waterbeemd, H., H., W.V.C., Eds.; 1995; Vol. 2.
81. StatSoft Inc., 2001.
82. Geer, L. Y.; Domrachev, M.; Lipman, D. J.; Bryant, S. H. *Genome Res.* **2002**, *12*, 1619.
83. Ursing, B. M.; van Enckevort, F. H.; Leunissen, J. A.; Siezen, R. J. *In Silico Biol.* **2002**, *2*, 1.
84. Thompson, J. D.; Higgins, D. G.; Gibson, T. J. *Nucleic Acids Res.* **1994**, *22*, 4673.
85. Opperdoes, F. R.; Szikora, J. P. *Mol. Biochem. Parasitol.* **2006**.
86. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J.Nucleic Acids Res. , *25*1997, 389.
87. Marchler-Bauer, A.; Bryant, S. H. *Nucleic Acids Res.* **2004**, *32*, W327.
88. Dea-Ayuela, M. A.; Bolás-Fernández, F. *Vet. Parasitol.* **2005**, *132*, 43.
89. Gharahdaghi, F.; Weinberg, C. R.; Meagher, D. A.; Imai, B. S.; Mische, S. M. *Electrophoresis* **1999**, *20*, 601.