

The use of generalized expression profiles for prediction of proteins associated with merozoite invasion

Galina V Glazko

Stowers Institute for Medical Research, 1000 E 50th St., Kansas City MO 64110

816-926-4455

gvg@stowers-institute.org

Arcady Mushegian

Stowers Institute for Medical Research, 1000 E 50th St., Kansas City MO 64110 and Department of Microbiology, University of Kansas Medical Center, Kansas City, KS 66160

816-926-4021

arm@stowers-institute.org

ABSTRACT

We propose a method to iteratively explore the expression profiles space in order to find profiles similar to the query. Starting from seven *Plasmodium falciparum* antigens as queries, we detected 596 similar expression profiles. Among those, 399 profiles were found by earlier distance-based approach, and 197 are newly detected profiles. The newly identified probes correspond to five known antigens and 108 other proteins. Many proteins in this latter category display properties compatible with a role in invasion, such as membrane localization, function in membrane remodeling and cytoskeleton dynamics, or enrichment in non-globular sequence domains. These same trends are also observed in the set of putative invasion-associated proteins identified by others.

Categories and Subject Descriptors

J.3 [Life and Medical Science]

General Terms

Algorithms, experimentation, validation.

Keywords

Microarray data analysis, profile similarity search, merozoite invasion proteins.

1. INTRODUCTION

The invasion of host cells by the malaria parasite *P. falciparum* is a multistep process involving (1) attachment of the merozoite to the red blood cell; (2) orientation of the apical end of the merozoite towards the blood cell membrane; (3) attachment of the apical end and formation of tight junction; and (4) interiorization process, sometimes accompanied by unexplained rotation of the merozoite [1]. Most of invasion-related proteins remain to be identified. Merozoite invasion proteins are promising vaccine candidates. We propose an approach for prediction of new invasion proteins on the basis of similarity of their expression profiles to those of the proteins with already established role in the invasion. In order to compare and match gene expression profiles, we propose the Ψ^2 algorithm. Each expression profile associated with a known invasion factor is used as a query in an iterative search of the expression profiles database, identifying all profiles that are similar to the query at a given level of significance.

1.1 Complex Expression Patterns

P. falciparum has developed a specialized mode of time-dependent transcriptional regulation, where maximal expression of genes involved in general cellular processes is followed by maximal

expression of parasite-specific genes [2]. Genes that are involved in invasion may be good candidates for vaccine development. Using genome-wide approach, Bozdech et al. (CAMDA 04 challenge data set, reference 2) looked for putative invasion-related genes. They studied the extreme of superimposed distributions of Euclidean distances between expression profiles of seven antigens and the rest of plasmodium transcriptome. The 5% of this distribution with the lowest distance was proposed as plausible vaccine candidates.

Not all gene expression in *Plasmodium* is phase-specific. For example, genes involved in antioxidant defense are not transcribed in-phase [3]. In fact, expression patterns of the known set of 28 genes involved in merozoite invasion [2] appear to be much less correlated than would be expected if they all had a phase-specific expression pattern. Indeed, we examined the distribution of correlation coefficients among expression profiles of these genes and found that it is not uniform ($\mu=0.6$, $\sigma=0.3$) and skewed to left (Fig. 1), towards non-significant correlations, even despite the occurrence of multiple, highly correlated probes for one transcript in this set. Thus, it may be prudent to separately explore the expression neighborhood of each invasion-related gene, without relying on correlations between several such genes.

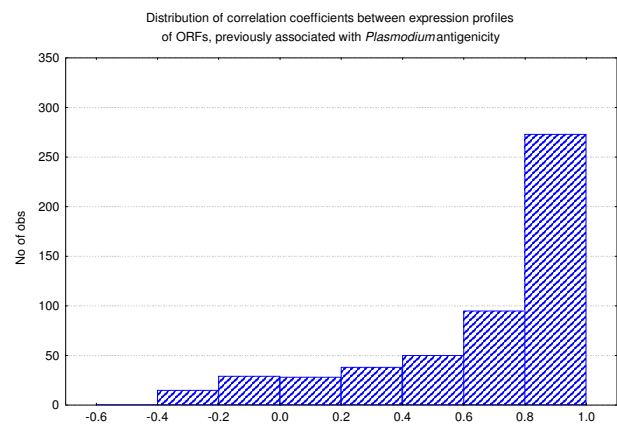


Figure 1.

2. AN ALGORITHM FOR SIMILARITY SEARCH IN THE EXPRESSION PROFILES SPACE

We developed Ψ^2 , an algorithm for similarity search in the expression profiles space. The logic of the algorithm is similar to

iterative sequence similarity search, namely, we would compare query profile with the expression profiles of all other genes, and construct condition-specific similarity matrix (CSSM) (conditions in this context are time points). CSSM matrix is updated at each iteration of the search, in order to account for new profiles exceeding the similarity threshold.

Before iteration, data are transformed into discrete vectors, as follows. For all profile in the expression profile set (EPS), the range of expression values, E_{\max} and E_{\min} , observed over all time points, is calculated. Then, the number of intervals or categories, K , is fixed and K intervals of expression, with $\text{step} = E_{\max} - E_{\min} / K$ are defined. Each expression profile is transformed into the discrete profile, where the i th expression value is replaced by the number of interval in which this expression value fall.

After discretization, the condition-specific scoring matrix (CSSM) for the entire data set is constructed. The rows correspond to different categories (intervals) and the columns correspond to the conditions (time points). For each category at every time point, we compute its frequency of appearance at this time point. For example, if there were no expression during first six hours, the first six columns will be filled with 0's over all categories except the first category, which corresponds to zero expression. CSSM matrix values for the entire data set, c_{ij} are constant and computed before all iterations.

Given a query (expression profile), our algorithm first extracts all expression profiles from the expression profiles set (EPS), which are significantly similar to the query profile. If the correlation coefficient between query profile, Q_i and EPS profile, P_i , $|\text{r}(Q_i, P_i)| > \text{SIGN}_{95\%}$, then (Q_i, P_i) form the High-Scoring Pair, HSP. The similarity measure, as well as its significance threshold ($\text{SIGN}_{95\%}$) are flexible. The algorithm proceeds through the following steps:

1. Find all HSPs.
2. Construct the weight matrix of the form $w_{ij} = \log(q_{ij}/c_{ij})$. Target frequencies q_{ij} for $\text{CSSM}_{\text{subset}}$ are computed from HSPs set, and background frequencies c_{ij} are computed from the entire data set.
3. Search the pattern space by computing scores between the $\text{CSSM}_{\text{subset}}$ and each profile; $S(\text{profile}) = \sum w_{ij}$, where i is run over all time points and j is the category observed in profile (from 1 to K). For these scores, construct the empirical distribution. Consider profiles with scores from a given percentile of the score distribution as new matches (we used 99%).
4. Add these profiles to the new matches list; update the $\text{CSSM}_{\text{subset}}$.
5. The process converges when we cannot find new profiles at step 3.

There are three parameters that be adjusted. The first parameter is the similarity threshold, which should be high, while providing a reasonably large set of HSPs to begin the iterations. The second parameter is the percentile of the score distribution that is used as the inclusion cutoff. Both parameters depend on the sample size. These are very similar to the cutoffs commonly used in sequence database search methods. The additional parameter that we use in this application is the number of categories, K . Discretization of expression profiles gives better resolution and sharper boundaries between groups of different profiles (unpublished observations).

3. EXPERIMENTAL RESULTS

3.1 Data Pre-Processing

We collected all probes (14) for seven best-known malaria vaccine candidates [2]. When one ORF was represented by multiple probes, we chose the probe with the highest correlation coefficient to other probes. The Quality Control data set [2] was used as the search space. Both missing data and outliers (deviating more than three s.d. from the mean for given profile) were replaced by the mean for this profile.

3.2 Parameters Tuning

Ψ^2 parameters were chosen to maximize the number of iterations (IT), maximize the ratio of new matches (Ψ^2 -U, i.e. unique matches from Ψ^2) to those obtained in the distance method (reference 2; abbreviated as 5%-ED in this work), and minimize the average number of matches found during all iterations. The interplay of these parameters and their effect upon the performance of the method are illustrated in Table 1.

Table 1. The average number of iterations, unique matches (detected by Ψ^2 alone), and matches found by both Ψ^2 and 5%-ED method.

CORR/K	IT	MATCHES	SHARED	Ψ^2 -U	5%ED-U
cor07_step05	1.1	508.3	418	605	1
cor07_step10	2.1	513.6	418	605	1
cor07_step15	1.7	509.4	418	605	1
cor07_step20	1.3	508.4	418	605	1
cor07_step25	1.1	508.3	418	605	1
cor07_step30	1.1	508.3	418	605	1
cor07_step35	1	508.1	418	605	1
cor07_step40	1	508.1	418	605	1
cor08_step05	1.7	339.7	418	356	1
cor08_step10	3.7	352.7	418	358	1
cor08_step15	2.1	339.1	418	356	1
cor08_step20	2.4	338.1	418	356	1
cor08_step25	1.3	334.9	418	356	1
cor08_step30	1.6	335.9	418	356	1
cor08_step35	1.4	336	418	356	1
cor08_step40	1.3	335	418	356	1
cor09_step05	3.3	193.1	408	125	11
cor09_step10	3.3	188.1	409	169	10
cor09_step15	4	196.4	409	187	10
cor09_step20	3.1	189.3	410	160	9
cor09_step25	2	179.3	408	152	11
cor09_step30	2.4	183	407	150	12
cor09_step35	1.9	180.6	407	164	12
cor09_step40	2	177.4	407	151	12

We fixed relatively high threshold for correlation between expression profiles in EPS and a query (0.9), to ensure that only highly correlated expression profiles are used to initialize the search. At this threshold the highest number of iterations (4), as well as the highest number of matches (187) found by Ψ^2 but

missed by 5%ED approach, was observed at K=15. We examined the results obtained with this parameter set in more detail.

3.3 Correlations Range During Iterations

Fig. 2 present the generalized expression profiles of seven query proteins, binned into some of 15 different categories. As expected, these expression profiles reach their maxima at the times corresponding to schizont stage [2].

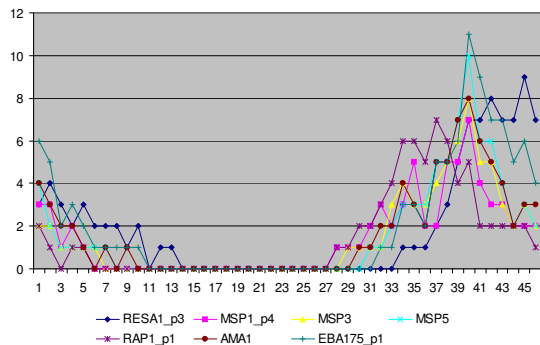


Figure 2

The first round of search collects only profiles with correlation of 0.9 or higher (Fig. 3, 45 generalized profiles found for query RESA1).

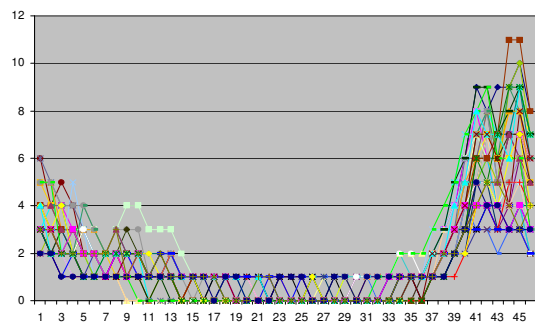


Figure 3

Further iterations bring together expression profiles with wider range of correlations (Fig. 4, query RESA1, average correlation is 0.678).

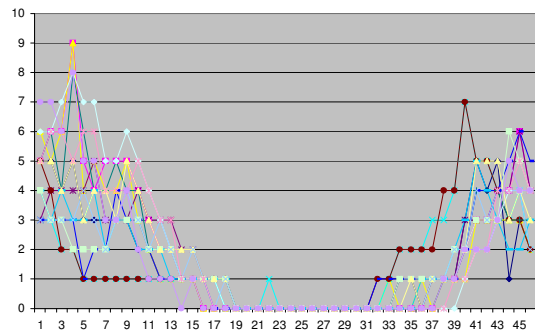


Figure 4

Overall, all ORFs found by Ψ^2 algorithm demonstrate clear trends of (1) gradual decrease of expression during ring stage, 1-12 hrs; (2) minimal or no expression during trophozoite stage, 12-30 hrs and (3) gradual increase of expression during schizont stage, 30-48 hrs (Figs. 3,4), similarly to what was observed with query antigen proteins (Fig. 2).

3.4 ORFs Found Only by Ψ^2 Algorithm

We analyzed in more detail the ORFs missed by the 5%-ED approach but scored as significantly similar to the queries in our approach. There were 187 such probes, corresponding to 151 unique ORFs (Supplementary Table 1). The fragment of the heat map for 10 of these proteins is shown in the Figure 5 (all heat maps can be found in Supplementary Figure 1, available upon request from the authors).

The average maximum time of expression for 151 ORFs corresponds to the beginning of the schizont stage (30 hours). Among them were several proteins with already known antigenic activity, such as RESA-H3, MSP8, octapeptide-repeat (ORA), PF70, membrane protein ag-1, RESA-2, tryptophan/threonine-rich antigen, and transmission-blocking target antigen. None of these proteins have been identified by 5%ED method. There were also 108 other proteins of unknown function.

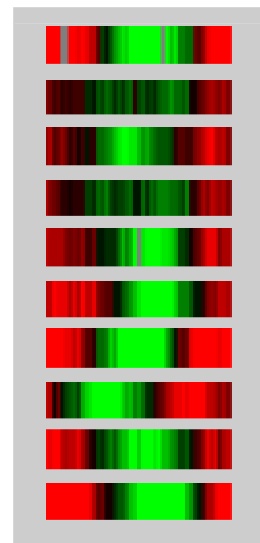


Figure 5

3.5 Sequence Analysis of 108 and 154 Merozoite Invasion Candidates

We compare two sets of hypothetical proteins (HP), those that were found by Ψ^2 approach only (HP_{s1} , 108), and by both Ψ^2 and 5%ED methods (HP_{s2} , 154). These two sets do not overlap, and it was interesting to see whether predicted biochemical properties of the proteins in both sets display any significant (similar or different) trends.

Table 2. Properties of proteins in two HPs set: HP_{s1} found only by Ψ^2 and HP_{s2} by both Ψ^2 and 5%ED approaches.

SET	Protein length	TM regions	%Loop regions	%Helical regions
HP_{s1}	888.39±8.33	1.17±0.02	91.08±0.13	8.92±0.13
HP_{s2}	929.08±6.85	1.28±0.02	94.09±0.07	5.85±0.07

The intrinsic properties of protein in two sets are nearly the same, except that HP_{s1} include slightly shorter proteins, than HP_{s2} and on the average its proteins contain more helical regions (Table 2; see Supplementary data available from the authors for description of prediction methods).

At the level of predicted molecular functions, the two groups of proteins are also quite similar (Supplementary table 2). Both sets are depleted of the housekeeping genes involved in genome

expression, in intermediate metabolism, and in signal transduction from cytoplasm to the nucleus. On the contrary, among the proteins with predicted enzymatic activity, there is a clear prevalence of domains involved in lipid metabolism synthesis and membrane remodeling (lipases, rhomboid-family intramembrane protease, MORN domains, ETRAMPs [4]). Also seen in both sets are proteins with chaperone activity (e.g. heavy metal chaperone, DnaJ domain), components of cytoskeleton and of secretory vesicles (actin-binding, microtubule associated proteins, dynamin family GTPase), and multiple protein kinases and phosphatases. Notably, the Ψ^2 and 5%ED methods recover different sets of proteins belonging to each of these functional categories. These observations are compatible with the idea of regulated changes in the cell surface and cell shape upon transitioning to the merozoite phase. Another telling property of both protein sets is the high proportion of transmembrane region in them [5]. Interestingly the bona fide antigen-related proteins, recovered in HP_{s1} and HP_{s2} are also different (RESA in the case of HP_{s1}) and AMA-1 and MSP7 in the case of HP_{s2}).

4. CONCLUSIONS

Our results suggest that the Ψ^2 approach is sensitive and specific. Starting with expression profiles of seven antigens, we iteratively searched the set of all expression profiles in Plasmodium, found previously unidentified similar profiles, and predicted molecular functions of many corresponding proteins. Although most of the protein sequences that we analyzed have not been implicated in merozoite invasion, the distribution of their properties seems to be

in good agreement with those of earlier proposed invasion candidates. The Ψ^2 approach may have broader applicability in matching binary and other vectors.

5. REFERENCES

- [1] Pasvol G. How many pathways for invasion of the red blood cell by the malaria parasite? *Trends Parasitol.*, 2003, 19, 430-432.
- [2] Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. The Transcriptome of the Intraerythrocytic Developmental Cycle of *Plasmodium falciparum*. *PLoS Biol.*, 2003, 1, E5.
- [3] Bozdech Z, Ginsburg H. Antioxidant defense in *Plasmodium falciparum*: data mining of the transcriptome. *Malar J.*, 2004, 3, 23.
- [4] Spielmann T, Ferguson DJ, Beck HP. etramps, a new *Plasmodium falciparum* gene family coding for developmentally regulated and highly charged membrane proteins located at the parasite-host cell interface. *Mol Biol Cell.*, 2003, 14, 1529-1544.
- [5] Di Cristina M, Spaccapelo R, Soldati D, Bistoni F, Crisanti A. Two conserved amino acid motifs mediate protein targeting to the micronemes of the apicomplexan parasite *Toxoplasma gondii*. *Mol Cell Biol.*, 2000, 20, 7332-7341.