HAL
open science

Détection de nouveaux domaines protéiques par co-occurence : Application à P. falciparum

Nicolas Terrapon, Olivier Gascuel, Laurent Brehelin

To cite this version:

Nicolas Terrapon, Olivier Gascuel, Laurent Brehelin. Détection de nouveaux domaines protéiques par co-occurence : Application à P. falciparum. JOBIM 2009 - 10es Journées Ouvertes en Biologie, Informatique et Mathématiques, Jun 2009, Nantes, France. pp.43-48. lirmm-00414954

HAL Id: lirmm-00414954

https://hal-lirmm.ccsd.cnrs.fr/lirmm-00414954v1

Submitted on 10 Sep 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Manuscrit auteur, publié dans "Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM), (2009)"

lirmm-00414954, version 1 - 10 Sep 2009

DÉTECTION DE NOUVEAUX DOMAINES PROTÉIQUES PAR CO-OCCURRENCE : APPLICATION À *P. falciparum*

Nicolas Terrapon[1 2], Olivier Gascuel[1], Laurent Bréhélin[1]

[1] LIRMM, Univ. Montpellier 2, CNRS, 161 rue Ada 34392 Montpellier Cedex 5 France
[2] CEA Grenoble iRTSV/LPCV, 17 rue des Martyrs, 38054 Grenoble cedex 9 France

Abstract: *Hidden Markov Models (HMMs) have proved to be powerful for protein domain identification. However, numerous domains may be missed in highly divergent proteins. This is the case for the proteins of Plasmodium falciparum, the main causal agent of human malaria. Here, we propose a method that uses domain co-occurrence to increase the sensitivity of the approach while controlling its false discovery rate. Applied to P. falciparum, our method identify (with an error rate below 20%) 482 new domains (versus 3482 in PlasmoDB), which involve 158 new GO annotations.*

Keywords: Hidden Markov Models, Protein Domains, Gene Ontology, Malaria.

1 Introduction

Les *modèles de Markov cachés* (HMM [1]) se sont révélés être un outil puissant pour l'identification de domaines protéiques grâce à leur capacité à capturer l'information spécifique à chaque position. Chaque HMM représente un domaine donné. Étant donné une nouvelle séquence protéique, l'approche probabiliste permet de calculer un score qui reflète la probabilité que le HMM ait généré la séquence. Ce score peut aussi être utilisé pour calculer une E-valeur, espérance du nombre de séquences ayant un aussi bon score dans une base de données de séquences aléatoires. La base de données en ligne Pfam (version 22.0) [2] propose une large collection de HMM modélisant des familles de domaines couvrant plus de 73% des protéines d'Uniprot[3]. Un certain nombre de domaines Pfam sont annotés dans la *Gene Ontology* ou GO [4]. L'annotation d'un domaine correspond aux informations communes à toutes les protéines ayant ce domaine [5], ce qui permet lorsqu'un nouveau domaine est identifié dans une protéine, de transférer les annotations GO du domaine à la protéine. De plus, Pfam fournit avec ses modèles des seuils, permettant d'affirmer la présence du domaine si le score de la séquence est supérieur au seuil. Cependant, chez certaines protéines fortement divergentes, cette approche n'est pas assez sensible pour permettre l'identification des domaines composants la protéine. Appliquée à *P. falciparum* par exemple (l'agent responsable de la forme létale de la malaria humaine), cette stratégie se révèle incapable de détecter le moindre domaine dans plus de 50% de ses protéines, tandis que de nombreux domaines semblent absents du répertoire de *P. falciparum* (seulement 1300 domaines distincts ont pu être identifiés). À titre de comparaison 2100 domaines sont répertoriés chez la levure, et concernent plus de 73% des protéines. Une des explications à ces difficultés réside dans le fort biais compositionnel des protéines de *P. falciparum*, induit par la composition à 80% de A+T de son génome. Relâcher les seuils requis pour la détection des domaines permettrait de plus nombreuses annotations, mais au prix d'un nombre d'erreurs important. Une solution est alors d'utiliser des informations supplémentaires pour filtrer parmi ces domaines

potentiels ceux qui ont le plus de chance d'être réellement présents. Dans cet article nous proposons d'utiliser la co-occurrence de domaines pour cela.

Les différentes études publiées concernant la combinatoire des compositions en domaines des protéines révèlent un certains nombre de propriétés. Les protéines composées des mêmes domaines ont généralement une fonction similaire [6]. La conservation de groupes de domaines au cours de l'évolution a été mise en évidence par plusieurs études montrant que le nombre de combinaisons de domaines identifiés dans la nature est infime en comparaison du nombre de combinaisons possibles : les domaines protéiques n'apparaissent qu'avec un nombre nombre limité d'autres domaines favoris au sein des protéines [7].

Nous présentons dans un premier temps notre méthode de recherche par co-occurrence ainsi qu'une procédure permettant de contrôler le taux d'erreur de la méthode. Nous validons ensuite notre approche grâce à des simulations sur la levure, puis nous présentons les résultats obtenus lorsqu'elle est appliquée à un organisme fortement biaisé comme *P. falciparum*.

2 Méthode

Nous proposons d'utiliser les propriétés de co-occurrence des domaines pour *certifier* la présence d'un domaine potentiellement présent dans une protéine à partir de la présence avérée d'un autre domaine. Notre approche consiste dans un premier temps à identifier parmi toutes les protéines de Uniprot, les paires de domaines montrant une co-occurrence forte (vérifiée par un test statistique) dans de nombreuses protéines. Ces paires de domaines conditionnellement dépendants (*PDCD*) forment alors une liste de référence qui est utilisée de la manière suivante. Considérons une protéine de notre organisme cible (par exemple *P. falciparum*) pour laquelle un ou plusieurs domaines sont déjà connus. En relâchant les seuils de score, les HMM de Pfam détectent un ou plusieurs nouveaux domaines potentiels. Si l'un de ces domaines forme, avec au moins un des domaines connus de la protéine, une paire faisant partie de la liste des *PDCD* de référence alors il est considéré comme certifié. Pour appliquer cette méthode de certification par co-occurrence, on a donc besoin de connaître, pour chaque protéine i de l'organisme étudié, l'ensemble de ses domaines *avérés* (A_i) et *potentiels* (P_i). Il faut aussi établir à l'aide de l'ensemble des protéines de composition connue, la liste de paires de domaines co-occurrents de référence, notée *PDCD* qui permet de certifier un domaine potentiel $x \in P_i$, grâce à un domaine avéré $y \in A_i$, si $(x, y) \in$ *PDCD*.

L'ensemble des domaines potentiels (P_i) se construit à partir des résultats de la recherche des HMM de Pfam sur la séquence protéique i grâce au logiciel `hmmer` [8]. Elle est paramétrée pour fournir l'ensemble des domaines dont l'E-valeur est inférieure à une valeur beaucoup moins stringente que la valeur seuil proposée par Pfam. Les résultats sont ensuite traités pour obtenir un ensemble de domaines non-recouvrants. Cette opération est effectuée grâce à un algorithme de pavage qui conserve en priorité les domaines possédant la meilleure E-valeur. À l'issue de cette phase, on conserve pour chaque protéine i l'ensemble des domaines potentiels non redondants P_i.

La base de connaissance des domaines avérés (A_i) peut être construite de différentes manières. La plus sûre est d'extraire directement des bases de données dédiées aux organismes, les domaines Pfam dont la présence a été certifiée par des experts, par exemple la base PlasmoDB [9] (version 5.5) pour *P. falciparum*. Elle peut aussi être obtenue en effectuant une recherche à l'aide des HMM de Pfam sur l'organisme cible en respectant les seuils proposés par Pfam. Cependant, d'autres bases de connaissance complémentaires peuvent être envisagées. On peut par exemple s'appuyer sur l'ensemble des domaines Interpro [5] répertoriés dans notre organisme cible (issus de PlasmoDB pour *P. falciparum*). L'utilisation de l'intégralité des bases de données d'Interpro permet alors de disposer d'informations

lirmm-00414954, version 1 - 10 Sep 2009

issues de 9 bases de domaines protéiques supplémentaires (SMART, PROSITE, Gene3D, Superfamily, PANTHER, Tigrfams, PRINTS, PIRSF, ProDom). En étendant de cette manière notre base de connaissances et en apprenant une liste de *PDCD* spécifique où chaque paire est composée d'un domaine Pfam et d'un domaine Interpro (non-Pfam), nous espérons pouvoir certifier plus de domaines, même dans des protéines où aucun domaine Pfam n'est connu. Néanmoins, comme pour la base Pfam, cette base limite la certification par co-occurrence à des protéines où au moins un domaine est déjà connu. Une autre base de connaissance complémentaire est de considérer les domaines potentiels (P_i) eux-mêmes comme base de connaissance. Dans cette solution, on essaye de certifier les domaines potentiels obtenus en relâchant les seuils de Pfam, par eux-mêmes (au risque d'un taux d'erreur plus important) afin de détecter des domaines Pfam dans des protéines où aucun domaine n'est connu.

La liste des paires de domaines conditionnellement dépendants est calculée à partir de l'ensemble des paires qui ont déjà été observées dans les protéines d'Uniprot chez d'autres organismes. Ces paires étant utilisées pour certifier la présence potentielle d'un domaine grâce à un autre domaine, elles doivent révéler une dépendance conditionnelle entre ces domaines, *i.e.* la présence de l'un doit être un indice fort de la présence de l'autre. Toutes les paires observées dans Uniprot ne satisfont pas ce critère. Par exemple, si deux domaines fréquents apparaissent avec de nombreux domaines différents (très versatiles), ils ne forment pas une paire conditionnellement dépendante. Tester la dépendance conditionnelle des paires de domaines revient à mesurer l'association de deux variables. On doit effectuer un test de comparaison entre deux proportions correspondant à l'observation simultanée de deux caractères différents sur les mêmes individus. Les individus sont les N protéines multidomaines d'Uniprot dont la composition en domaines est connue. Les deux caractères observés dans ces protéines sont la présence (ou l'absence) des domaines formant chaque paire. Une solution à ce problème peut être apportée par un test de corrélation de type χ^2. Nous avons choisi d'appliquer un test exact de Fisher, plus adapté pour de petits échantillons comme c'est le cas ici. Pour chaque paire de domaines une P-valeur peut donc être calculée. Si cette P-valeur est inférieure à un certain seuil (typiquement 5%) l'hypothèse nulle est rejetée, les domaines sont considérés comme conditionnellement dépendants, et la paire est ajoutée à la liste des *PDCD*.

Contrôle du taux de faux positifs : À partir des domaines potentiels, des domaines avérés et de la liste des *PDCD*, on est capable de certifier un certain nombre de domaines inédits. Une question est alors de pouvoir estimer le nombre de domaines certifiés par erreur par notre approche. Pour cela nous proposons d'estimer l'espérance du nombre de nouveaux domaines que notre approche certifierait sous l'hypothèse H_0 où tous les domaines potentiels étaient prédits de manière aléatoire. Cela peut être réalisé par simulation, à l'aide d'une procédure de permutation aléatoire des différents domaines potentiels des protéines. Permuter les différents domaines crée une situation dans laquelle les domaines potentiels sont indépendants des domaines avérés, tout en préservant la distribution de ces domaines, ainsi que la distribution du nombre de domaines potentiels et avérés par protéine. La procédure de permutation est la suivante. Dans un premier temps, l'ensemble des domaines avérés associés aux protéines est fixé. Puis on collecte l'ensemble des domaines potentiels de toutes les protéines, et on les redistribue aléatoirement à travers les différentes protéines en créant de nouveaux ensembles de domaines potentiels P_i^* de même taille que les ensembles P_i originaux. On applique ensuite notre méthode sur ces domaines potentiels, et on comptabilise le nombre de domaines potentiels qu'elle certifie. On réitère cette procédure un grand nombre de fois (typiquement 1000), et on moyenne les résultats. Ce nombre moyen de domaines certifiés sous l'hypothèse H_0 est comparé au nombre de certifications réalisées sur les données originales. Le taux de faux positifs (estimation du *False Discovery Rate*, ou *FDR*) de la méthode est estimé par le ratio :

$$FDR = \frac{\text{espérance du nombre de certification sous } H_0}{\text{nombre de domaines certifiés sur les données originales}}.$$

lirmm-00414954, version 1 - 10 Sep 2009

lirmm-00414954, version 1 - 10 Sep 2009

En jouant sur le seuil d'E-valeur utilisé pour définir les domaines potentiels, on peut donc, grâce à cette procédure, contrôler le *FDR* associé à nos prédictions.

3 Résultats

La première expérience réalisée consistait à nous assurer de la capacité de la méthode à trouver les domaines qui échappent aux seuils de Pfam à cause d'une dérive trop importante des séquences protéiques. Le principe est le suivant. Les HMM de Pfam sont utilisés avec leurs seuils de score pour déterminer l'ensemble des domaines de référence chez *S. cerevisiae*, organisme choisi pour la qualité de ses annotations. On fait ensuite subir aux séquences de la levure une évolution rapide vers la composition de *P. falciparum* à l'aide du programme `seqgen` [10]. Nous avons ainsi créé 4 jeux de séquences protéiques artificiels de divergence croissante (grâce à des taux t de substitution de 0.1, 0.25, 0.5 et 0.75, une matrice de substitution, *WAG*, et une distribution d'acides aminés cible : la distribution moyenne chez *P. falciparum*), sur lesquelles on applique la procédure suivante. Dans un premier temps, chaque HMM est utilisé avec les seuils de Pfam pour détecter les domaines présents. On s'attend à ce qu'un certain nombre de domaines de référence ne soient plus détectés à cause de la dérive des séquences. Dans un second temps, nous relâchons les seuils (à une E-valeur de 10) et appliquons la méthode de certification par co-occurrence en utilisant les domaines Pfam encore détectés par les seuils comme base de connaissance. On espère ainsi retrouver une partie des domaines précédemment perdus. Les résultats sont présentés dans le tableau 1. Par exemple, pour $t = 0.5$, des 907 domaines perdus, 645 sont potentiellement retrouvable (*i.e.* sont présents dans une protéine pour laquelle au moins un autre domaine est encore détecté), et 491 sont effectivement retrouvés. De plus, 60 inédits (absents des domaines de référence) sont également détectés. Pour les taux de substitution élevés, on remarque que la proportion d'inédits parmi les domaines certifiés (*i.e.* $\frac{\text{Domaines inédits}}{\text{Domaines retrouvés} + \text{Domaines inédits}}$) est proche du taux d'erreur estimé par notre procédure ce qui tend à valider cette procédure. Pour les taux bas, par contre, on remarque que le taux d'inédits est sensiblement plus haut que le taux d'erreur estimé. Une question est alors de savoir si parmi ces inédits une certaine partie ne serait pas de "vrais" domaines non encore référencés chez la levure. Pour vérifier cette hypothèse, nous avons calculé parmi les domaines retrouvés qui possèdent une annotation GO, la proportion possédant une annotation non référencée chez la protéine (dernière colonne du tableau). On constate que la proportion de domaines ayant une annotation GO inédite est beaucoup plus basse que la proportion de domaines inédits, et plus proche de notre *FDR*. Les autres domaines (apportant des annotations déjà connues) qui constituent l'essentiel des domaines inédits, sont concordants avec les annotations connues de la protéine. Cela semble indiquer que les "vrais" inedits (apportant des annotations GO inédites) sont en effet rare, comme on peut s'y attendre chez la levure, et donc qu'une partie des domaines inédits ne sont pas des faux positifs mais des domaines réellement présents que nous certifions grâce à notre approche.

Nous avons ensuite appliqué notre méthode à *P. falciparum* en utilisant les trois sources d'information détaillées en introduction : les domaines Pfam connus, les domaines Interpro non-Pfam déjà connus, et les domaines potentiels eux-mêmes. La figure 1 présente les résultats obtenus pour différents seuils d'E-valeurs (en abscisse), en utilisant les domaines Pfam référencés dans PlasmoDB [9] et en utilisant une liste de PDCD selectionnée avec une P-valeur seuil de 5%. On constate comme attendu que le nombre de domaines certifiés ainsi que le *FDR* augmentent avec le seuil d'E-valeur utilisé pour la sélection des domaines potentiels. On peut donc, suivant que l'on désire un plus grand nombre de domaines certifiés ou un FDR faible, jouer sur le seuil d'E-valeur pour générer un ensemble de prédictions en accord avec l'objectif privilégié.

lirmm-00414954, version 1 - 10 Sep 2009

Taux substitution	Domaines de référence	Domaines perdus	Potentiellement retrouvable	Domaines retrouvés	Domaines inédits	*FDR* Estimé	Proportion nvx GO
0.1	2407	149	145	134	274	11.5%	15%
0.25	2407	346	301	265	171	9.2%	7.8%
0.5	2407	907	645	491	60	5.4%	3.1%
0.75	2407	1436	747	501	12	4%	0.3%

Tableau 1. Résultats sur la levure après évolution. ”Taux substitution” indique le taux de divergence des séquences, ”Domaines de référence” les domaines des protéines multidomaines de la levure originale, ”Domaines perdus” correspond aux domaines non retrouvés par les seuils de Pfam sur les séquences divergentes, ”Domaines retrouvés” les domaines perdus que l’on retrouve par notre méthode de certification, ”Domaines inédits” le nombre de domaines inédits à l’ensemble de référence trouvé en plus par notre méthode, et ”Proportion nvx GO” la proportion de domaines ayant une annotation GO inédite vis à vis de la protéine.

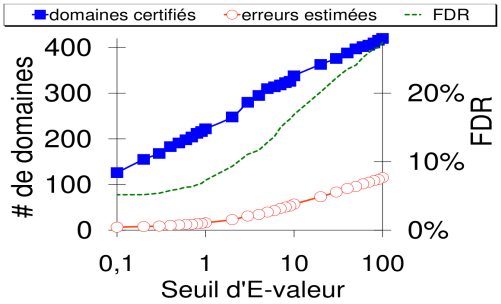

Fig. 1. Évolution du nombre de certifications, de l’estimation du nombre d’erreurs et du *FDR* en fonction de l’E-valeur (en abscisse). Le nombre de domaines certifiés (carrés) et le nombre d’erreurs estimées (cercles) évoluent en ordonnées sur l’axe de gauche, et le *FDR* (en pointillés) sur l’axe de droite.

Le tableau 2 présente l’ensemble des résultats obtenus en utilisant les trois bases de connaissances pour différentes tranches de *FDR* : les prédictions ayant un FDR inférieur à 10%, celles ayant un FDR compris entre 10 et 25%, et le cumul de ces deux ensembles, qui correspond à un FDR global inférieur à 20%. Par exemple, pour un *FDR* inférieur à 20%, 482 nouveaux domaines sont certifiés, parmi lesquels 361 correspondent à l’identification d’une nouvelle famille de domaines Interpro dans la protéine. Ils représentent un apport de 13,8% de domaines par rapport à l’ensemble des 3482 domaines Pfam connus chez *P. falciparum* d’après PlasmoDB. Les domaines certifiés l’ont été grâce aux domaines Pfam connus pour 351 d’entre eux, des domaines Interpro pour 253 et des domaines potentiels pour 110 (avec un certain recouvrement, certains domaines étant certifiés par 2 ou 3 de ces bases). De plus, ces domaines certifiés avec un *FDR* inférieur à 20% ont permis la découverte de 126 types de domaines qui n’avaient jamais été observés dans une protéine de *P. falciparum* auparavant. Ces domaines vont s’ajouter au 1421 types de domaines connus chez *P. falciparum* (*cf.* section 1), soit une amélioration d’environ 9%. Enfin, parmi les nouveaux domaines certifiés chez *P. falciparum*, un certain nombre possèdent des annotations GO inédites qui peuvent être transférées aux protéines. Par exemple pour un *FDR* inférieur à 20%, les domaines certifiés apportent un total de 158 nouvelles annotations GO chez *P. falciparum* (soit 2% d’annotations supplémentaires si l’on se rapporte aux 8312 annotations GO de *P. falciparum*), 107 provenant d’un nouveau domaine ayant été certifiés par co-occurrence avec des domaines Pfam connus, 58 avec des domaines Interpro connus et 36 avec les domaines potentiels. Par exemple nous avons identifié une protéine impliquée dans la synthèse de la *cobalamine* (vitamine B12), une molécule nécessaire au développement de *P. falciparum* [11], ce qui constitue donc une cible thérapeutique potentielle.

FDR	<10%				10%< . . . <25%				Cumul <20%			
Base avérés	Pfam	Interp.	Pot.	Toutes	Pfam	Interp.	Pot.	Toutes	Pfam	Interp.	Pot.	Toutes
Domaines certifiés	224	131	52	284	127	122	58	198	351	253	110	482
Nvlles Familles Interpro	155	97	45	205	100	90	44	156	250	187	89	361
Domaines inédits chez *Pf*	59	33	23	76	36	27	12	50	95	60	35	126
Nvlles annotations GO	64	25	11	81	43	33	25	77	107	58	36	158

Tableau 2. Tableau récapitulatif des résultats sur *P. falciparum* pour différents tranches de *FDR*. ”Base avérés” correspond aux bases de connaissance des domaines avérés utilisées pour la certification : ”Pfam”, ”Interp.” pour Interpro, ”Pot.” pour les domaines potentiels et ”Toutes” pour les résultats cumulés des trois bases. ”Nvlles annotations GO” indique le nombre de nouveaux termes GO transférés aux protéines.

4 Conclusion

Nous avons présenté une méthode améliorant la sensibilité de la détection de domaines protéiques par des modèles probabilistes, en s’appuyant sur les propriétés de co-occurrence des domaines. Cette méthode qui a été initialement développé pour l’étude d’organismes dont l’annotation est pauvre (dûe à un protéome à fort biais compositionnel), peut aussi s’appliquer à des organismes déjà bien annotés. Nos résultats montrent qu’elle permet de certifier un nombre important de domaines, tout en contrôlant le taux d’erreur en fonction de l’objectif privilégié (nombreux nouveaux domaines ou *FDR* stringent). Appliquée à *P. falciparum*, elle permet par exemple de certifier 482 nouveaux domaines avec un *FDR* inférieur à 20% et d’apporter 158 nouvelles annotations GO à ses protéines.

Remerciements

Ce travail est soutenu par le projet ANR PlasmoExplore (ANR-06-MDCA-014). Nous remercions tout particulièrement Éric Maréchal, ainsi que l’ensemble des membres du projet PlasmoExplore.

References

[1] R. Durbin, S. Eddy, A. Krogh and G. Mitchison, *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*, Cambridge University Press, New York, 1998.
[2] R.D. Finn, J. Tate, J. Mistry, P.C. Coggill, S.J. Sammut, H.R. Hotz, G. Ceric, K. Forslund, S.R. Eddy, E.L.L. Sonnhammer and A. Bateman, The Pfam Protein Families Database. *NAR*, 36:D281-D288, 2008.
[3] The UniProt Consortium, The Universal Protein Resource (UniProt). *NAR*, 36:D190-D195, 2008.
[4] The Gene Ontology Consortium, The Gene Ontology (GO) project in 2006. *NAR*, 34(Database issue):D322-D326, 2006.
[5] N. Mulder, R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, D. Binns, *et al.*, New developments in the InterPro database. *Nucleic Acid Research*, 35(Database Issue):D224-228, 2003.
[6] M. Gerstein and H. Hegyi, Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Research*, 11:1632-1640, 2001.
[7] I. Cohen-Gihon, R. Nussinov and R. Sharan, Comprehensive analysis of co-occuring domain sets in yeast proteins, *BMC Genomics*, 8:161, 2007.
[8] S.R. Eddy, Profile Hidden Markov Models. *Bioinformatics*, 14:755-763, 1998.
[9] A. Bahl, B. Brunk, J. Crabtree, D. Gupta, J.C. Kissinger, D.S. Pearson, D.S. Roos DS, J. Schug, C.J. Jr Stoeckert *et al.*, PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data. *NAR*, 31(1):212-215, 2003.
[10] A. Rambaut and N.C. Grassly, Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, 13:235-238, 1997.
[11] S.M. Chemaly, C.T. Chen and R.L. van Zyl, Naturally occurring cobalamins have antimalarial activity. *J Inorg Biochem.*, 101(5):764-773, 2007.

lirmm-00414954, version 1 - 10 Sep 2009