PROCEEDINGS OF
THE 2009 INTERNATIONAL CONFERENCE ON
BIOINFORMATICS & COMPUTATIONAL BIOLOGY

BIOCOMP 09

Volume I

Editors

Hamid R. Arabnia
University of Georgia, USA

Mary Qu Yang
National Institutes of Health, USA

Associate Editors

Youping Deng, Chien-Tsai Liu, Ashu M. G. Solo
Yanqing Zhang

WORLDCOMP'09
July 13-16, 2009
Las Vegas Nevada, USA
www.world-academy-of-science.org

Identifying the Active Site of Ribonucleoside Hydrolase of E. Coli Encoded by RihC

A. Farone[1], M. Farone[1], A. Khaliq[2], P. Kline[3], T. Quinn[2] and Z. Sinkala[2]
[1]Department of Biology, Middle Tennessee State University , Murfreesboro, Tennessee, USA
[2]Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, Tennessee, USA
[3]Department of Chemistry, Middle Tennessee State University, Murfreesboro, Tennessee, USA

Abstract— *We predict the potential active and catalytic sites, the transition state and how it is stabilized, and the mechanism of Ribonucleoside hydrolase of E. Coli(rihC). Our approach is based on well-known primary sequence analysis techniques. A canonically associated extreme value distribution is used to assess the significance of the prediction. Parameters for the extreme value distribution are computed directly from data. Our new approach reproduces well known examples in the literature, and yields the construction for the general case. In addition to our results for rihC, our work makes use of strategies that could be useful for a range of medically significant scenarios.*

Keywords: Extreme value distribution, sequence analysis, random walk, score function, ribonucleoside hydrolase, rihC

1. Introduction

We report on recent progress toward identifying the active site within the primary sequence, for ribonucleoside hydrolase rihC of Escherichia coli. Using sequence analysis of rihC and inosine-uridine nucleoside hydrolase (IU-NH) from C. fasciculata, a number of amino acids have been identified as potentially important in the interaction of the enzyme with its substrate. Sequence analysis of IU-NH from C. fasciculata (the most studied nucleoside hydrolase) and rihC identifies (likely) amino acids involved in the active site of the enzyme. We also obtain: a list of potential active site residues; information on the mechanism of the enzyme encoded by rihC; and identification of the transition state. From our findings we propose the following about rihC: His233 in its primary sequence acts as a proton donor to activate the uridine leaving group. The catalytic site contains Asp10, Asn165, and His233. The active site contains Asp10, Asp14, Asp15 and Asp234 along with Ile121, coordinate of divalent cation. Its mechanism is that His233 acts as an acid to protonate the N7 of the purine or N3 of the pyrimidine of the leaving group if the base is cytosine. The $Ca2^+$ ion together with Asp10 activates a water molecule which nucleophilically attack ribose. Asp10 accepts a proton from the water molecule. This leads to a transition state which spontaneously dissociates to form ribose and purine or ribose and pyrimidine. We also propose that the transition state is stabilized by Asn165.

The death rate reported for malaria, trypanosomiasis, and other infections caused by protozoan parasites exceeds one million per year [7]. This impact on human health shows the importance of developing improved treatments for these diseases. The nucleoside hydrolase rihC is part of the nucleic acid alternative pathway of E. coli and Salmonella enterica serovar Typhimurium which catalyzes the hydrolysis of different nucleosides to ribose and the corresponding base. Nucleoside hydrolases catalyze the reaction of

$$\text{Nucleoside} + H_2O \longrightarrow \text{ribose} + \text{purine or pyrimidine.} \quad (1)$$

This reaction is vital for the salvage pathways of protozoan parasites and also is used by bacteria.

Nucleoside hydrolases, while vital in the nucleoside salvage pathway of protozoans, are apparently absent in mammals. "NH's are widely distributed in nature and have been found in bacteria, yeast, protozoa, insects and mesozoa. ... The metabolic role of the NH's is well established only for parasitic protozoa (Trypznosoma, Keishmania, giardia and so on.) .... Parasitic protozoa rely on the purine salvage pathway for survival because - in contrast to other living organisms - they lack a *de novo* biosynthetic pathway for purines. Considering this divergence in purine metabolism between parasite and host, the parasitic NHs have been studied extensively in recent years as potential targets for chemotherapeutic intervention." [11] In other words, because mammals lack these enzymes, specific differences in the nucleic acid pathways between mammals and protozoans have made the nucleoside hydrolases the target for possible development of chemotherapeutic agents. Indeed, insight into the structure and mechanism of the involved enzymes could help in the development of inhibitors and other anti-parasitic drugs.

Escherichia coli has multiple pathways for the salvage of nucleosides. One of these pathways consists of a group of hydrolases capable of breaking down nucleosides to ribose and the corresponding base. E. coli has three different genes for these hydrolases, one of which is rihC, and is capable of hydrolyzing both purines and pyrimidine ribonucleosides(1).

The nucleoside hydrolases studied to date have been grouped into three classes [11]. They are (1) relatively base nonspecific nucleoside hydrolases exemplified by the IU- NH isolated from C. fasciculata; (2) purine-specific

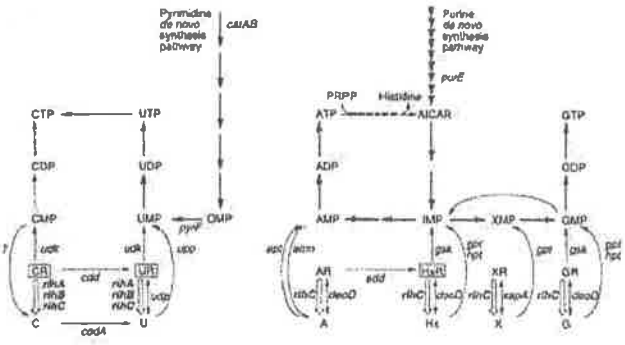

Fig. 1: Pathways of pyrimidine and purine metabolism in E. coli. rihC is shown to hydrolyze multiple substrates and is the only hydrolase shown to act on purines .

nucleoside hydrolases exemplified by the IAG-NH isolated from Trypanosoma brucei brucei and Trypanosoma vivax; and (3) 6-oxopurine-specific inosine–guanosine preferring nucleoside hydrolases exemplified by GI-NH isolated from C. fasciculata. The enzyme being investigated in the present study is encoded by rihC, and belongs to class 1. A well known enzyme of the class IU-NH comes from C. fasciculata, has been extensively studied and has been partially characterized by X-ray crystallography and kinetic isotope effects using inosine as the labeled substrate [6], [5].

Additional results for IU-NH from C. fasciculata can be found in [3], [2], [5], [10]. The enzyme catalyses the hydrolysis of all purine and pyramidine nucleosides, the hydrolysis forms ribose and the associate base. The enzyme has a preference for inosine and uridine as substrates. His241 in its primary sequence acts as a proton donor to activate the hypoxathine leaving group. The catalytic site contains Asp10, Asn168, and His241. The active site contains Asp10, Asp14, Asp15 and Asp 242 along with Thr126, coordinate of divalent cation (possibly calcium). Its mechanism is that His241 acts as an acid to protonate the N7 of the leaving purine. The $Ca2^{+}$ ion together with Asp10 activates a water molecule which nucleophilically attack ribose. Asp10 accepts a proton from the water molecule. This leads to a transition state which spontaneously dissociates to form ribose and purine. The transition state is stabilized by Asn168.

We would like to know more about rihC, including such things as enzyme preferences, substrate specificity, kinetic constants, enzyme mechanism, conformational structure, location and topology of active sites within the three dimensional conformations. (See, for example [Inosine-uridine nucleoside hydrolase (IU-NH) from crithidia fasciculata]; [3]; [2]; [5]; and [10].) At this time, however, little is known about the locations in the primary sequence for the active sites of rihC.

2. Our approach

Our approach makes use of sequence alignment methodology [9], [4], [8]. We use the primary sequence of IU-NH from C. fasciculata as a basis for comparison with rihC. Residuals central to the activity of an enzyme are conserved in the enzyme family. The method uses scoring functions to assess local similarity between the rihC sequence and IU-NH of C. fasciculata sequence. We obtain a BLOSUMr matrix in a new way, so that the choice of $r\%$ is intrinsically tied to the data base. A more detailed description of our basic mathematical results follows.

Dayhoff's method for comparing closely related species does not work well when aligning evolutionary divergent sequences [1]. Sequence changes over long evolutionary time scales are not well approximated by compounding small changes that occur over short time scales. The BLOSUM (BLOCK substitution matrix) matrices help address this problem. Henikoff and Henikoff constructed these matrices using multiple alignments of evolutionary divergent proteins. The probabilities used in the matrix calculation are computed by looking at "blocks" of conserved sequences found in multiple protein alignments. These conserved sequences are assumed to be of functional importance within related proteins. To reduce bias from closely related sequences, segments in a block with a sequence identity above a certain threshold are clustered giving weight 1 to each such cluster (Henikoff and Henikoff). For the BLOSUM62 matrix, this threshold is set at 62%. In other words, pairs are counted only between segments less than 62% identical. One uses a higher numbered BLOSUMr matrix for aligning two closely related sequences and a lower number r for more divergent sequences.

To choose a BLOSUMr matrix, practitioners tend to invoke various hypotheses on evolutionary distances between sequences of interest. In alignment applications such as BLAST, commonly used defaults are BLOSUM62 and BLOSUM50 [8]. The BLOSUM62 matrix does an excellent job detecting similarities in distant sequences. Another approach is to align the sequences by using several BLOSUM matrices [12]. A natural question though is what natural threshold to use. Is there a systematic way to choose $r\%$ for the BLOSUMr matrix? In our new approach, we have have developed a method for selecting an $r\%$ (and so a BLOSUMr matrix) that is intrinsically tied to the data base, and does not require *a priori* data on evolutionary distances.

In sequence analysis, an alignment of two sequences and a score function S determine a random walk. Typically, a scale is chosen so that the score function takes on values approximated by integers, with step sizes from the set

$$T = \{-c, -c+1, ..., 0, 1, 2, ..d - 1, d\}, \qquad (2)$$

with $c > 0, d > 0$ both strictly positive integers. The cumulative score between two sequences is the random walk.

Note that, as in the general theory [4], [8], we assume

$$p_{-c} > 0,\ p_d > 0,\ \sum_{j=-c}^{d} jp_j < 0 \tag{3}$$

and that the greatest common divisor of all positive integers $j \in T$ for which $p_j > 0$ is equal to unity. To assess similarity that could be of biological significance, one investigates the distribution of high scoring segments. The method therefore appeals to extreme value distributions, asymptotic to the form

$$P(S > S_0) \approx 1 - \exp\left(-K^* \exp\left(-\lambda^* S_0\right)\right). \tag{4}$$

In applications, one must determine the parameters K^* and λ^*. Mathematically λ^* is the unique solution to an equation of the form

$$\phi(\lambda) = 1, \tag{5}$$

where $\phi(\lambda)$ is a convex function satisfying

$$\phi(0) = 1,\ \phi'(0) < 0,\ \phi(\lambda) > 0,\ \phi''(\lambda) > 0 \tag{6}$$

for all $\lambda > 0$ and

$$\phi(\infty) = \infty. \tag{7}$$

Computationally it is straightforward to obtain numerical solutions for λ^*. Obtaining the parameter K^*, however, is rather more involved [4]

The parameter K^* can be expressed in terms of A/C and λ^* [4]. The two terms A and C correspond to quantitative features of the random walk space and can be given in terms of probability functions R_{-j} [8] and Q_k respectively, where $j = 1, 2, ..., c$ and $k = 1, 2, ..., d$. It is then possible to approximate K^* by making use of a convergent series

$$\Psi\left(R_{-1}, R_{-2}, ..., R_{-c}, Q_1, Q_2, ..., Q_d\right) \tag{8}$$

that appears as an exponent [9].

We have developed two new and direct approaches for computing the parameter K^*. One approach is primarily computational and the other is analytic. Our numerical work aligns well with known results. The computational strategy is partly motivated by keeping mathematical terms closely aligned with biological construction (and therefore accessible to experiment). This leads to an efficient algorithm that allows for direct approximation of the probability functions

$$R_{-1}, R_{-2}, ..., R_{-c}, Q_1, Q_2, ..., Q_d \tag{9}$$

which by equation (8) consequently produces an approximation for the parameter K^*.

3. Conclusions

Our computational approach reveals the possibility of obtaining an analytic system of equations for the probability functions

$$R_{-1}, R_{-2}, ..., R_{-c} \text{ and } Q_1, Q_2, ..., Q_d. \tag{10}$$

In the literature, a closed form for R_{-j} is known only for elementary cases where there are at most three terms

$$R_{-1}, R_{-2}, R_{-3}. \tag{11}$$

See for example [9]; [4]; and [8]. Our new approach reproduces these known examples, and yields the construction for the general case. In addition to our results for rihC highlighted in Section 1, our work makes use of strategies that could be useful for a range of medically significant scenarios.

$$T = \{-c, -c+1, ..., 0, 1, 2, ..d-1, d\}. \tag{12}$$

References

[1] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, "A model of evolutionary change in proteins," *Atlas of Protein Sequence and Structure*, vol. 5, pp. 345-352, 1978.

[2] M. Degano, S. C. Almo, J. C. Sacchettini, V. L. Schramm, "Trypanosomal nucleoside. A novel mechanism from the structure with a transition state inhibitor. N-ribohydrolase from Crithidia fasciculata." *Biochemistry* , vol. 37, pp. 6277-6285, 1998.

[3] M. Degano, D. N. Gopaul, G. Scapin, V. L. Schramm, J. C, "Sacchettini Three-dimensional structure of the inosine-uridine nucleoside N-ribohydrolase from Crithidia fasciculata." *Biochemistry* , vol. 35, pp. 5971-5981, 1996.

[4] W. J. Ewens and and G. R. Grant, *Statistical Methods in Bioinformatics*, 2nd ed., New York: Springer-Verlag, 2004.

[5] D. N. Gopaul, S. L. Meyer, M. Degano, J. C. Sacchettini, V. L. Schramm, "Inosine-uridine nucleoside hydrolase from Crithidia fasciculata. Genetic characterization, crystallization, and identification of histidine 241 as a catalytic site residue." *Biochemistry* , vol. 35, pp. 5963-5970, 1996.

[6] B. A. Horenstein, D. W. Parkin, B. Estupinan, V. L. Schramm, "Transition-State analysis of nucleoside hydrolase from crithidia fasciculata," *Biochem.* vol. 30, pp. 10788-10795, 1991.

[7] C. Hunt, N. Gillani, A, Farone, and P. C. Kline," Kinetic isotope effects of nucleoside hydrolase from Escherichia coli," *Biochimica et Biophysica Acta*, vol. 1751, pp. 140-149, 2005.

[8] A. Isaev, *Introduction to Mathematical Methods in Bioinformatics*, Berlin, German: Springer-Verlag, 2004.

[9] S. Karlin, and S. F. Altschul, "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes," *Proc. Natl Acad. Sci. USA*, Vol. 87, pp. 2264-2268, 1990.

[10] D. Mazumder, K. Kahn, and T. C. Bruice, "Computer simulations of trypanosomal nucleoside hydrolase: determination of the protonation state of the bound transition-state analogue," *J. Am. Chem. Soc.*, vol. 124, pp. 8825-8833, 2003.

[11] W. Versees and J. Steyaert, "Catalysis by Nucleoside hydrolases,"*Current Opinion in Structural Biology*, vol. 13, pp.731-738, 2003.

[12] K. H. Zimmermann, *An introduction to protein informatics*, Boston USA: Kluwer Academic publishers, 2003.