Recruitment of rare 3-grams at functional sites: Is this a mechanism for increasing enzyme specificity?
Dror Tobi and Ivet Bahar*

Address: Department of Computational Biology, School of Medicine, University of Pittsburgh, Pittsburgh PA 15261, USA
Email: Dror Tobi - drt6@pitt.edu; Ivet Bahar* - bahar@ccbb.pitt.edu
* Corresponding author

Abstract

Background: A wealth of unannotated and functionally unknown protein sequences has accumulated in recent years with rapid progresses in sequence genomics, giving rise to ever increasing demands for developing methods to efficiently assess functional sites. Sequence and structure conservations have traditionally been the major criteria adopted in various algorithms to identify functional sites. Here, we focus on the distributions of the 20³ different types of 3-grams (or triplets of sequentially contiguous amino acid) in the entire space of sequences accumulated to date in the UniProt database, and focus in particular on the rare 3-grams distinguished by their high entropy-based information content.

Results: Comparison of the UniProt distributions with those observed near/at the active sites on a non-redundant dataset of 59 enzyme/ligand complexes shows that the active sites preferentially recruit 3-grams distinguished by their low frequency in the UniProt. Three cases, Src kinase, hemoglobin, and tyrosyl-tRNA synthetase, are discussed in details to illustrate the biological significance of the results.

Conclusion: The results suggest that recruitment of rare 3-grams may be an efficient mechanism for increasing specificity at functional sites. Rareness/scarcity emerges as a feature that may assist in identifying key sites for proteins function, providing information complementary to that derived from sequence alignments. In addition it provides us (for the first time) with a means of identifying potentially functional sites from sequence information alone, when sequence conservation properties are not available.

Background

Proteins perform a variety of biological functions, the efficiency and specificity of which are determined to a large extent by their intrinsic sequence-structure properties. Many biophysical and biochemical activities such as binding specific substrates, catalysis, or allosteric responses involve particular sequence motifs at functional sites. Identification of functional sites, or understanding sequence-to-function mapping, in general, has been a major goal in computational molecular biology and bioinformatics. With rapid accumulation of genome scale sequence data there is an ever increasing need for efficient assessment of potential functional sites.

Among different criteria adopted to identify functional sites, sequence conservation has probably been the most widely used, based on the fact that functional residues are often conserved across all or the majority of members in a
given family of proteins. Algorithms developed for detecting conservation patterns are traditionally based on multiple sequence alignments [1-6]. Other algorithms use a combination of sequence and structure comparisons provided that structural information is available, which usually yield relatively higher predictive abilities [7-11]. A set of criteria was proposed by Thornton and coworkers for detecting catalytic residues, which was successfully implemented in an algorithm for identifying active sites [9]. Other strategies include pattern matching such as TESS [12], FFF [13] and SPASM [14] that locate functional sites by detecting small three-dimensional motifs. Proteins dynamics has been shown to be another property that can be advantageous examined to detect functional sites [15]. We have recently shown that the global hinge centers of enzymes co-localize with their catalytic sites, pointing to the importance of coupling between mechanics and chemistry for enzymatic activity [16]. In addition, analyzing patterns of communication and shortest paths in protein structures modeled as networks has proven useful in identifying allosteric sites [17,18].

In the present study, we propose another property that can serve as a criterion for detecting functionally important sites: scarcity or rareness in the space of sequence motifs. Motifs composed of three sequential amino acids (3-grams) are considered here as the shortest, yet distinctive, sequence fragments that can provide statistically significant information. The choice of 3-grams in the following work is based on preliminary studies as well as our previous work [19] that showed that 1-grams and 2-grams do not confer enough specificity while there is no major difference in the performance of 3-grams and 4-grams (due to the fact that some strong signals characteristic of 3-grams may be overlooked upon examination of 4-grams), while 3-grams lend themselves to more accurate statistics. A given 3-gram will be termed ‘rare’ if its probabilistic occurrence in the UniProt (Universal Protein Resource) [20] is significantly lower than that expected from an unbiased distribution, as will be explained in quantitative details below.

A notable feature emerging from this study is the propensity of proteins’ active sites to populate rare sequence motif, akin to the notion that rare structural motifs co-localize with functionally important sites [21-23]. This selectivity suggests that enzymes tend to recruit such distinctive/rare sequences at their active sites to increase their specificity, consistent with the higher information content associated with rare events. We illustrate the functional importance of such rare 3-grams in three proteins, C-Src kinase, hemoglobin, and tyrosyl-tRNA synthetase.

Results and discussion

Distributions of 3-grams

A given protein sequence of N residues is viewed as a collection of N – n + 1 overlapping words of n letters, termed n-grams (also called n-tuples), composed of n contiguous amino acids along the sequence. For each protein, we consider all 3-grams (n = 3) or triplets, by sliding a window of three amino acids along the sequence, thus leading to a total of (N – 2) triplets. The natural frequency of a given 3-gram of type i (1 ≤ i ≤ 203) is defined as

\[
p_{i}^{\text{uni}} = \frac{C_{i}}{\sum C_{j}}
\]

where \(C_{i}\) is the number of occurrence in the UniProt, and the summation is performed over all 203 types of 3-grams in the 172,233 sequences compiled as of March 1, 2005, which contain a total number \(C_{\text{tot}} = 62,256,868\) of 3-grams. Each of the 203 distinctive types is represented in UniProt. Their counts vary in the range 170 < \(C_{i}\) < 67,264, with the lower and upper bounds corresponding to the 3-grams MCW and AAA, respectively. Figure 1 displays the distribution of the counts \(C_{i}\). The inset shows a closer view of the percentage of 3-grams types whose UniProt counts fall in successive grids of size \(\Delta C_{i} = 500\), for the range \(C_{i} \leq 1.5 \times 10^{4}\). We note that less than 1% of the 203 types have

![Figure 1](http://www.biomedcentral.com/1471-2105/8/226)

**Figure 1**

**Distribution of 3-grams in UniProt.** Histogram of the counts \(C_{i}\) for all types (1 ≤ i ≤ 203) of 3-grams, shown for grids of size \(\Delta C_{i} = 10^{3}\). The mean \(<C_{i}\>\) and standard deviation \(\sigma_{C_{i}}\) are 7,822 and 6,682, respectively. The inset shows the portion of the curve for \(C_{i} \leq 15,000\) in more detail. The ordinate is the percentage of 3-grams within intervals of \(\Delta C_{i} = 500\). The peak (6.07%) occurs at the interval 2500 ≤ \(C_{i}\) ≤ 3000. 3-grams that are one standard deviation away from the mean towards lower counts are termed ‘rare’ 3-grams. Their counts are lower than \(<C_{i}> - \sigma_{C_{i}} = 1,140\). There is a total of 480 such 3-grams (i.e. 6% of all the 8,000 types of 3-grams), and their cumulative frequency of occurrence evaluated from the ratio of their total count to \(C_{\text{tot}}\) is 0.595%.
UniProt counts $C_i$ lower than 500, and 14.97% has $C_i < 2,000$. Rare 3-grams are defined as those in the lower end of the histogram, located one standard deviation away from the mean. They comprise 6.0% of all 3-gram counts in the UniProt (see Figure 1 caption).

The unbiased probability of a given 3-gram is

$$p_i^\text{uni} = \frac{1}{203^3} = 0.000125,$$

down to 4% of $C_i$ or $C_{tot}/203^3 = 7782$ in the UniProt. The 3-grams whose count is lower than this value by a factor of $f = 2$ compose 10% of all the counts $C_{uni}$ in UniProt, and those lower by $f = 3$ amount to 4% of $C_{tot}$.

**Scarcity scores and enhancement factors**

The frequency $p_i^\text{uni}$ given by eq (1) will also be termed the 'observed' or 'natural' probability. We will associate with each 3-gram type a scarcity score given by

$$s_i^\text{uni} = -\ln[p_i^\text{uni}] \quad (2)$$

which provides a measure of the information content each 3-gram carries. The distribution of scarcity scores is given in Figure 2. They vary in the range $6.83 < s_i^\text{uni} < 12.81$. Panel A displays the scores sorted in descending order, plotted against 3-gram type index, and panel B displays their histogram (distribution among the 203 types).

**Figure 2**

**Scarcity Scores.** (A) Scarcity scores (eq 2) plotted as a function of 3-gram index, sorted in descending order. (B) Histogram of scarcity scores. The abscissa shows the number of 3-grams having scarcity scores values lying in successive ranges of size 0.25 shown along the ordinate. The peak occurs at $s_i^\text{uni} = 9.0 \pm 0.125$. Unique 3-grams (Figure 1) lie in the range $s_i^\text{uni} > 10.908$, or $i < 480$, indicated by the arrow, and their observed cumulative frequency is $p_i^\text{uni} < 1.83 \times 10^{-5}$.

based on grids of size $\Delta s_i^\text{uni} = 0.25$. The arrows display the threshold adopted for defining rare 3-grams.

The 'expected' probability of 3-gram $i = \{XYZ\}$ on the other hand, based on the natural occurrences of individual amino acids, assuming the probabilistic occurrence of amino acids to be independent of their sequential neighbors, is

$$p_i^\exp = p^0_X p^0_Y p^0_Z \quad (3)$$

where $p^0$ is the natural frequency of amino acid of type X (for details see Table S1 in Additional file 1). The enrichment in a given 3-gram count will be described by the enhancement factor

$$a_i = \frac{p_i^\text{uni}}{p_i^\exp} \quad (4)$$

Figure 3A compares the observed and expected counts of 3-grams. While the slope gives an average enhancement factor $<a_i> = 1.02$ with a correlation coefficient of 0.96, a more detailed examination of expected and observed 3-grams Figure 3C shows that the two distributions exhibit statistically significant departure ($\chi^2$, $P < 0.001$). The expected histogram is more broadly distributed than that observed. Some 3-grams that show substantial departures from their expected frequencies (i.e. enhancement factor $a_i \neq 1$) are labeled in panel B. Among over-represented rare 3-grams we observe CCC, WWW, WYW, CWC, CHC and WWW. We also note that AAA and LLL are distinguished by their high $p_i^\text{uni}$, and HHH and QQQ by their enhancement $a_i \geq 6$. The HHH enrichment may, however, arise from tagged histidines, rather than reflecting a natural preference.

The present approach permits us to assign a scarcity score to each amino acid in a given sequence without recourse to sequence alignment. In the calculations below we have performed similar analyses for protein families, where the occurrence of rare 3-grams at specific locations where examined with regard to their conservation properties. Given an alignment of homologues sequences, an average scarcity score was assigned to each position $j$ along a sequence $k$, based on the minimum value,

$$p_i^\text{uni} |_{jk} = \min\{p_i^\text{uni} |_{a_k}, p_i^\text{uni} |_{b_k}, p_i^\text{uni} |_{c_k}\}$$

of the three consecutive 3-grams of types $(a, b, c)$ that contain residue $j$, using
where the summation is performed over the homologous sequences \( 1 \leq k \leq m \) in the alignment.

Proteins tend recruit rare 3-grams at their active sites

The biological significance of rare 3-grams has been examined by analyzing the results with regard to the dataset of 59 non-redundant enzyme/ligand complexes compiled by Gutteridge and Thornton [24], shortly referred to as the GT dataset (Table 1). Two non-overlapping subsets of 3-grams were retrieved from the GT dataset. The first is composed of all 3-grams located at the active sites, and the second includes all other residues. The two subsets comprise 3,514 and 17,450 3-grams, respectively. Active site 3-grams are defined as those containing at least one heavy atom located within 5Å from the ligand.

The UniProt frequencies of these two groups of 3-grams yielded the histograms displayed in Figure 4A. To obtain these histograms, the ensemble of 3-grams in the respective subsets 1 and 2 of active site residues (gray) and other residues (black) have been organized into groups of different UniProt frequencies, and the counts of 3-grams in each group have been determined. The abscissa shows the ranges of UniProt counts, divided into intervals of \( \Delta C_i = 1000 \), starting from \( C_i \leq 1,000 \) (labeled 1000), and the ordinate represents the population (or probability) of each group. Not all 203 types of 3-grams were represented in this set of 59 enzymes.

Comparison of the two histograms immediately reveals the tendency of the active site 3-grams to populate the lower UniProt counts. For example, let us consider the rare 3-grams (with \( C_i < 1,140 \) or \( s_j^{uni} > 10,908 \); see Methods and Figure 2). 11.67% of active site 3-grams fall in this range, whereas the percentage drops to 4.35% when 3-grams at other regions are considered. Therefore, rare 3-grams are over-represented in the active sites by a factor of

\[
< s_j^{uni} > = -\ln[< p_j^{uni} >] = -\ln[m^{-1}\sum_k p_j^{uni}] |_{jk} \tag{5}
\]

where the summation is performed over the homologous sequences \( 1 \leq k \leq m \) in the alignment.

### Table 1: GT (Gutteridge-Thornton, 2005)24 dataset of 59 enzymes*

| PDB Code | Chain | PDB Code | Chain | PDB Code | Chain | PDB Code | Chain |
|----------|-------|----------|-------|----------|-------|----------|-------|
| 1cib     |       | 1kbqAC   |       | 6alDA    |       | 1n2uA    |       | 1o7n    |       | 3apr    |       |
| 8acn     |       | 1d6A     |       | 1d6A     |       | 1eexABC  |       | 1pnf    |       | 6ceB    |       |
| 1a16     |       | 3daA     |       | 1f5A     |       | 1jcr     |       | 2dx     |       | 1d2A8    |       |
| 1b66A    |       | 1b57     |       | 1jcr     |       | 1eixAB   |       | 1ibv    |       | 4reqAB   |       |
| 1b74     |       | 1e7y     |       | 1gai     |       | 1eixAB   |       | 1ibv    |       | 4reqAB   |       |
| 1m7yA    |       | 1dz      |       | 1gff     |       | 1lee     |       | 1q1A    |       | 2sc      |       |
| 1rruA    |       | 1k9sAD   |       | 1gm7AB   |       | 1erm     |       | 1e8G     |       | 1oneA    |       |
| 1dag     |       | 1ue      |       | 1gkm     |       | 1hduA    |       | 1mt     |       | 1ra2     |       |
| 1hm2A    |       | 1esd     |       | 1o6G     |       | 1ma0     |       | 1bo0    |       | 1ldm     |       |
| 1lesw    |       | 1dud     |       | 1e9gA    |       | 1mka     |       | 1g49B   |       |          |       |

* The table lists the PDB code and chains used. Care was taken not to double count 3-grams in cases where the active site is located at the interface of two identical subunits.
Next, we examined the identity of 3-grams at the active sites, and focused in particular on the rare 3-grams that are recruited. Given that not all of the 20\(^3\) types of 3-grams were represented in the GT dataset, not all types of 3-grams might be either seen at the active sites. We first analyzed the amino acid composition of active sites. For each amino acid, the probabilities of being located at the active site versus elsewhere were calculated. The ratio between these two probabilities, termed the active site propensity, is shown in Figure 5A. Four amino acids, Cys, His, Gly, and Trp, are distinguished by their high active site propensities. This is in agreement with previous work [25] where His and Cys were shown to have the highest propensity for serving as catalytic residues via their side chains, while Gly has the highest propensity of catalyzing reactions via its main chain. We also note that, in the present study, Arg and Glu exhibit relatively low tendencies to locate at active sites, while these two amino acids were reported by Bartlett et al [25] to have relatively high catalytic propensities. This difference presumably arises from the different definitions adopted for active sites in the two studies: we refer to all amino acids located in the neighborhood of the ligand, while Bartlett et al [25] referred to residues directly involved in catalytic activity.

Cys and His, two residues that have the highest active site propensity (both in terms of their ligand-binding and catalytic activities) are among the least frequent amino acids in terms of their natural occurrence. Figure 5A shows the relation between the frequencies of amino acids in UniProt and their active site propensities. While the data are scattered, there is a discernible anti-correlation between the two sets of data, consistent with Figure 4; rare amino acids have a higher tendency to be in active sites. Clearly this preference becomes more pronounced when 3-grams, as opposed to monomers, are examined, as can be seen from the skewed distribution of the frequency of active site residues towards lower counts in Figure 4.

Rare 3-grams may arise from three possibilities: (1) these may be words in which the individual letters have low natural frequencies themselves (i.e. \(a_i \approx 1\)), (2) these may be rare ‘words’, i.e. rare combinations of particular amino acids (\(a_i << 1\)); or (3) these may be rare despite their enhancement (\(a_i >> 1\)) due to the extremely low natural frequencies of the individual amino acids that compose them. Counterparts of Figure 3 plotted for the active site 3-grams and other 3-grams showed that both groups fall in the first category (\(a_i \approx 1\)). Their observed counts plotted against the expected ones yield a slope of 1.075 ± 0.008. The rare 3-grams belonging to the two groups, on the other hand, exhibited a slight enhancement of \(a_i = 1.11\) and 1.07, for active site and other residues, respectively. Thus, 3-grams recruited at active sites are usually combinations of residues which already have relatively low nat-
ural frequencies of occurrence. Yet, particular 3-grams exhibit significant departure from their expected occurrence probabilities, e.g. SSS, GGG, RRR, GSG (over-represented), and LEL (under-represented). See Table S2 and S3 in Additional files for details. Among rare 3-grams at active sites, also exhibiting a large enhancement in the UniProt (case 3, above), we distinguish WWH and HCW.

**Scarce 3-grams in the GT dataset**

Scarcity scores were calculated for all 3-grams in each of the 59 proteins in the GT dataset. The postulate that rare 3-grams (above a threshold scarcity score) preferentially locate in active sites was examined for a series of threshold scores. For each enzyme, the rates of true positives (TP) and false positives (FP) were evaluated as a function of threshold scarcity scores, TP rates being defined as the fraction of active site 3-grams that exhibit scarcity score above the threshold value. Table 2 shows the average ratios between the rates of TPs and FPs for seven different thresholds. The uppermost threshold ($s_{\text{max}} = 11$) shows more than threefold enhancement in the rate of TPs relative to FPs, while the discriminative power of the scarcity score vanishes as we lower the threshold. A detailed list of TP and FP rates along with the top ranking 3-grams are provided for each enzyme in the Additional files Table S4.

**Illustrative results for three cases**

The functional role of rare 3-grams at proteins’ functional sites is illustrated here for three test cases, a Src tyrosine kinase, hemoglobin and a tyrosyl-tRNA synthetase. For each case, a set of homologous proteins has been compiled and aligned, and the scarcity score $s_j$ of amino acid(s) at each position has been calculated. The analysis reveals that stretches of rare amino acids are involved in functional roles even beyond ligand binding sites.

**Src kinase**

The Src family of non-receptor tyrosine kinases is a member of the tyrosine kinases superfamily and plays an important role in cell differentiation, proliferation and survival [26]. C-Src, a member of Src protein kinase family, contains two peptide binding modules (SH3 and SH2 domains), a catalytic tyrosine kinase domain (composed of N- and C-lobes) and a C-terminal regulatory tail (for a review see [27]) (Figure 6A). Its phosphorylation at activation loop residue Y527 results in intramolecular interaction between the SH2 domain and the phosphorylated C-terminal tail, triggering a transition into an inactive conformation with the SH2 and SH3 domains moving away from the catalytic site clamping the latter in a strained conformation. Dephosphorylation of Y527 and phosphorylation of Y416, on the other hand, activate the kinase via movement of the activation loop and opening of the catalytic cleft between the N- and the C- lobes of the kinase domain.

Figure 6B shows the scarcity scores $s_j$ as a function of residue number along the c-Src sequence, based on UniProt counts of 3-grams, averaged over the set of $m$ aligned homologous sequences (eq 5). Six n-grams (3 $\leq$ n $\leq$ 6) composed of contiguous 3-grams are distinguished by their high scores, D$_{117}$WWL$_{120}$, E$_{146}$EWYF$_{150}$, Y$_{382}$VH$_{384}$, I$_{426}$KWTA$_{430}$, D$_{444}$VWSF$_{448}$, and C$_{496}$QCWRK$_{501}$, in addition to individual residues H319, K200, C478 and P485.

**Table 2: Results for the GT dataset**

| Scarcity Score Threshold | TP/FP* |
|--------------------------|--------|
| 11.0                     | 3.152  |
| 10.5                     | 1.936  |
| 10.0                     | 1.543  |
| 9.5                      | 1.432  |
| 9.0                      | 1.353  |
| 8.5                      | 1.139  |
| 8.0                      | 1.041  |

* The average ratio between true positives (TP) rates and false positives (FP) rates for the dataset of 59 enzymes

---

**Figure 5**

**Active site propensity of amino acids and comparison with their natural frequencies.** (A) For each type of amino acids, the frequencies of occurrence in the active and non-active sites were calculated. The ratio between them is defined as the active site propensity (ordinate). (B) Natural (UniProt) frequency of occurrence vs. active site propensity for each type of amino acid. Less frequent amino acids tend to have a higher active site propensity, and vice versa, in accord with the histograms in Figure 4. This effect becomes more pronounced at the level of 3-grams.
D_{117}WW_{120} is located in the SH3 domain, and interacts with the linker that connects the SH2 and kinase domains when the kinase adopts its inactive conformation (Figure 6A). The SH3-linker interactions have been pointed out to serve as an independent mode of regulation for Hck, a Src family member [28]. In addition, Trp118, at the ligand-binding site of SH3 [29], was found to be essential to the stability of the SH3-VSL12 complex [30].

The second stretch of amino acids, E_{146}EWYF_{150}, coincides with the SH2-SH3 linker which was shown by Kuriyan and coworkers [31] to work as an inducible "snap-lock" that clamps the SH2 and SH3 domains upon phosphorylation of Y572. Mutations of residues in the linker have been shown to lead to a constitutive activation of c-Src [31].

The third stretch, Y_{382}VH_{384}, forms the Src catalytic loop. Y_{382} takes part in key interactions that stabilize the kinase domain in the inactive conformation [32] and H_{384} belongs to the conserved HRD motif of the catalytic loop. Lys200 at SH2 domain interface interacts with the C-terminal regulatory tail.

The other three stretches of amino acids I_{426}KWTA_{430}, D_{444}VWSF_{448}, and C_{496}QCWRK_{501} form a cluster located at and beneath the substrate binding region (Figure 6A). The portion Q_{497}CWR_{500} is composed of two overlapping 3-grams that have the highest scarcity scores in the protein. Notably, the central residues in these stretches, W499, C498, W428, W446, are highly or fully (bold) conserved, lending support to their critical role.

A systematic analysis of amino acid conservation conducted for c-Src protein family members demonstrated that amino acids distinguished by their high scarcity scores tend to be conserved among the family members (see Conclusion), despite their being rare in the UniProt. Conserved sites distinguished by their scarcity include, in addition to the four residues listed above, W148, K427, T429, F448, S447, A430, H319, H384, C487, D444 and V445. See Additional files Table S5 for details.

Figure 6C shows the two motifs derived for Src kinase from PROSITE database [33]. The first is the stretch of twenty two amino acids L_{273}GQGCFG ... GTTRVAIK_{295} corresponding to the ATP binding region and the second is the stretch Y_{382}VHRDLRAANILV_{394} corresponding to the catalytic loop. Except for the 3-gram Y_{382}VH_{384} identified in both panels A and C, the 3-grams deduced from scarcity score and PROSITE analyses appear to provide complementary information.
Hemoglobin

Hemoglobin (Hb) is a tetramer composed of two α- and two β-subunits (α1, α2, β1, and β2) organized in two dimers, α1β1 and α2β2, symmetrically positioned around a central water-filled cavity (Figure 7A). Each subunit has a heme that binds oxygen. The oxygenation of Hb is cooperative, i.e., binding of a first O2 enhances the O2 affinity of the other subunits. Following the MWC model [34], Hb exists in two conformations in rapid equilibrium: the T state with low affinity for oxygen binding and the R state with high affinity. The intrinsic oxygen binding affinity of the tetramer is nearly 300-fold lower than that of its free α1β1 dimer [35], implying that the interactions at the dimer-dimer interface interfere with substrate binding unless relieved by a conformational switch.

The <S>mi</S> profiles for subunits α and β are presented in Figure 7B and 7C, respectively. Five stretches of residues are distinguished by their high scarcity scores in subunit α: α A12AWGK16, α R31MF33, α Y42FPHF47, α H87AH89, and α S102HCL105, colored magenta in panel A. In subunit β, β W15GK17, β Y35PWTQ39, β L91HCD94 and β H143KYH146 (cyan) emerge as rare sequences.

Notably, all of these stretches of amino acids distinguished by their high scarcity score assume functional roles in Hb. α R31MF33 and α Y42FPHFD47 are directly involved in binding the heme group. The mutant R31S results in abnormal hemoglobin – Hb Prato [36]. α Tyr42 forms an inter-subunit hydrogen bond with β Asp99 across the dimer-dimer interface in the T state of hemoglobin but not on the R state. Kavanaugh et al [37] showed that this hydrogen bond and other interactions associated with the side chain of α Tyr42 make a major contribution to the stability of the T state. α H87AH89 corresponds to the heme binding site, with α His87 coordinating the Fe atom. The Fe-Nε bonds between the heme group and α His87 or β His92 are essential for the cooperativity of hemoglobin [38]. α His103 is located inside a central cavity lined with excess of positively charged ionizable groups. It has been suggested [39] and experimentally validated [40] that the mutual repulsion of these ionizable groups increases the oxygen affinity by raising the free energy of the T state. Using network analysis, del Sol et al confirmed that α His103 plays an important role in allosteric communication [17]. We also note the critical position of α His103 at the α1β1 interface with α Ser102 pointing toward the heme binding pocket. Among the residues participating in rare sequences in subunit β, in addition to the heme-binding β H92, we note β K17, the substitution of which by Asn has been reported to lead to abnormal Hb J Amiens [41]. The segment β Y35PWTQ39 is located at the convergence of the α1β1 and α2β2 interfaces.
and is important for allostery. Mutations of $\beta$ 37W (e.g. $\beta$ W37E and $\beta$ W37G) have been shown to lead to tetrameric Hb with functional properties similar to those of the $\alpha\beta$ dimers; that is, high oxygen affinity with no cooperativity [42,43]. Finally, the 4-gram $\beta$H143KYH146 at the C-terminal end of subunit $\beta$ and in particular H143 is a binding site for allosteric effectors [44,45].

In addition to the above experimental evidence, we note that using the Gaussian network model we have shown that two segments, $\alpha$ F36-H45 and $\beta$ T87-$\beta$ N102, play a key role in the allosteric transition of Hb from constrained (T form) to flexible (or relaxed) (R2 form) state [46]; these two regions include the so-called switch region [47]. The first stretch partially overlaps with $\alpha$ Y42FPHF46 and the second contains $\beta$ L91HCD94. These results lend further support to the critical role of the rare $n$-grams, and in particular $\beta$ His92, in propagating allosteric signals. Finally, we note that out of 37 rare residues identified in Hb, 14 are fully conserved (entropy $S = 0$), and 17 highly conserved ($S < 0.75$; see Additional files Table S6 for more details).

Upon scanning the hemoglobin sequence against the PROSITE database (Figure 7D) four histidines are retrieved, two ($\alpha$ His58 and $\alpha$ His87) on $\alpha$-subunits, and two ($\beta$ His63 and $\beta$ His92) on $\beta$-subunits. As mentioned above, $\alpha$ His87 and $\beta$ His92, also detected among the rare 3-grams, coordinate the Fe atom of the heme group, while $\alpha$ His58 and $\beta$ His63 are located at the same region across the heme plane.

Tyrosyl-tRNA synthetase

Tyrosyl-tRNA synthetase catalyzes the attachment of tyrosine to its cognate tRNA after activation of the tyrosine via formation of tyrosyl adenylate. Its structure has been solved in complexed form with an inhibitor (Figure 8A) that mimics the binding of the activated tyrosine (tyrosyl adenylate) in the active site [48]. The calculated $<s^\text{uni}>$ profile shown in Figure 8B reveals three $n$-grams distinguished by their uniqueness: $Y_{124}DWIG_{128}$, $D_{194}QWGN_{198}$, and $Y_{252}QFW_{255}$ (Figure 8A, magenta). The former two are located at the substrate binding site. The latter forms contacts with the conserved motif HIGH that comprises the catalytic His45 [49] and with the amino acids F231, G232, and T234 of the K_{230}FGKT_{234} mobile loop (Figure 8A, cyan). This loop was found to stabilize the transition state in the activation reaction of tyrosyl by ATP [50]. We note that W126, D194, Q195, W196, G197, N198, Y252 and Q253 are fully (bold) or highly conserved. See Table S7 in the Additional files for more details. PROSITE scanning reveals the

\[P_{39}TADSLHIGHL_{49}\] stretch of amino acids at the active site, which includes the catalytic His45 (Figure 8C).

**Figure 8**

(A) Identification of tyrosyl tRNA synthetase unique $n$-grams. Tyrosyl-tRNA synthetase (PDB:3TS1 [48]) is shown (white) in complex with tyrosyl adenylate intermediate (orange). The pair H45I46 (cyan) of the catalytic His, the residues K_{230}FGKT_{234} of the mobile loop (cyan) and the three most unique $n$-grams (magenta) are shown as sticks. (B) Scarcity score of amino acids. 24 homologous (E value < 10^{-57}) of *Bacillus stearothermophilus* were used to calculate the average uniqueness score. Three unique $n$-grams (filled circles) are distinguished. (C) PROSITE motifs. The loop containing the catalytic His at the binding site (magenta) is identified using PROSITE.

**Results in the absence of sequence alignment**

In the above three examples the scarcity score was calculated by averaging $p_{jk}^m$ over a set of homologous proteins. However, in principle, scarcity score calculations can be equally performed for a single sequence as the
ingredient is the UniProt frequencies for triplets along the sequence. The alignments are simply used to magnify (increase the accuracy of) the signals that could, otherwise, be extracted from single sequence analysis. As shown in the figures S1–S3 in the Additional files, the \( s_{ij}^{uni} \) profiles based on single sequences closely approximate those derived from multiple sequence alignments. For hemoglobin subunit \( \alpha \), for example, six stretches of amino acids are distinguished by their high scarcity scores: \( W_{14}GK_{16} \), \( R_{31}MF_{33} \), \( F_{43}PHF_{46} \), \( D_{75}MPN_{78} \), \( H_{87}AH_{89} \), \( S_{102}HCL_{105} \), in accord with the results presented above. The only stretch of amino acids that differs from the above analysis is \( D_{23}MPN_{26} \). Thus, when sequence conservation algorithms fail, the 3-gram analysis of single chains can be advantageously resorted to for identifying potentially functional sites.

**Conclusion**

The present study shows that relatively rare sequence motifs are preferentially recruited at enzymes’ active sites. The probabilistic occurrence of rare 3-grams at active sites is higher than that of other 3-grams by a factor of more than 2.6 and higher that of exposed 3-grams by a factor of 5.4 according to the results obtained for the GT dataset of 59 enzymes (Figure 4). Detailed analysis of each enzyme in this dataset demonstrated that a scarcity score threshold of 11.0, permits us to identify 3-grams located at/near active sites, with a TP rate more than three times larger than the FP rate. These results were obtained from the sequence of the individual enzymes without recourse to sequence alignment or any amino acid conservation information.

The preferential selection of rare 3-grams at functional sites may not be a property of enzymes, exclusively, but that of proteins, in general, as the current application to hemoglobin suggests. We have illustrated and discussed here the functional importance of rare 3-gram motifs for three proteins: c-Src, hemoglobin and tyrosyl-tRNA synthetase.

As a further assessment of the functional importance of rare \( n \)-grams, we examined the conservation of these \( n \)-grams among the members of the three families of proteins on which we focused. Figure 9 displays the results, reported in terms of the Shannon entropies [51], for two groups of residues: (i) the rare residues distinguished by their high scarcity scores (above the threshold values indicated in Figures 6B, 7B, C and 8B), and (ii) all other residues. The respective numbers of residues in the two sets are 83 and 1240. The histograms (percentages) are displayed for bins of size 0.25. We note that unique residues exhibit a stronger tendency to be conserved (despite their low counts in the UniProt) among the members of the examined families, lending further support to their functional importance (see Additional files for details).

While our analysis is purely sequence-based, it appears to be conceptually in accord with two structure-based studies that indicate that rare structural motifs correlate with proteins functional sites. Petock et al [22] showed that 9-residue fragments of rare backbone conformations often form parts of ligand-binding sites, protein-protein interactions, and domain-domain contacts. Likewise, Novotny and Kleywegt [23] found that, even though left-handed helices are rare, when they do occur, they are structurally or functionally significant. The present analysis shows that not only structural motifs, but sequence motifs as well, tend to be highly distinctive at active sites. This selection of rare sequence motifs seems to be driven by specificity requirements. Notably, the rarest amino acids (Cys, His, Trp and Met; Figure 5B) have unique functional properties: Cys and Met are sulfur-containing amino acids and are sensitive to oxidation. Likewise, His is unique as its protonation depends on pH, and Trp is distinguished by its bulky size and significant contribution to stabilization of hydrophobic cores or interfaces. It has been suggested that proteins may even sustain a decrease in their stability in order to form effective functional sites [21,22]. In accord with the above notion, these highly specific residues, Trp, His, Met, and Cys, may be used by enzymes for precise functional purposes, their specificity being amplified in 3-grams where two or more of these amino acids are juxtaposed.
Identification of proteins' functional sites using computational tools is a major challenge in computational biology in the post-genomic era [52,53]. Sequence and structure conservation are the major criteria used by current algorithms to identify functional important sites. Identification of functional sites for orphan proteins (for which no homologous protein exists) is beyond the scope of algorithms that use conservation as major criterion for identifying these sites. Notably, we repeated the scarcity score analysis by using only a single chain and we were still able to identify functionally important sites (see Additional files), demonstrating that the potentially functional sites of orphan proteins can also be assessed by our methodology, and that conservation and scarcity are two different features. We note that the average scarcity score of scarce and conserved residues will be higher than that of scarce and non-conserved residues, as the scarcity score is evaluated as an average over all aligned sequences. However, calculations repeated for single query sequences, as opposed to those averaged over multiple aligned sequences showed that such effects are negligibly small and the profile of scarcity scores for examined proteins are robust. In addition, we have compared the scarcity scores (based on UniProt frequencies) and the Shannon entropies at the corresponding amino acid positions. Computations for the three test cases revealed that residues participating in rare 3-grams (Tables S5–S7) are not necessarily conserved or vice versa. However, they do exhibit a higher tendency to be conserved compared to other amino acids, as illustrated in Figure 9. Therefore, it is conceivable that a combined approach exploiting both conservation and scarcity may enhance the score/signal associated with particular sites, thus increasing the ability to accurately identify functional motifs.

**Methods**

**Datasets**

Present calculations are based on the data extracted from UniProt [20] version 4.2. This database was chosen because it is manually curated with minimal level of redundancy and still comprehensive. Furthermore, UniProt contains entries from multiple organisms and it is therefore less susceptible to biases in 3-gram distributions of specific species as noted by Ganpathiraju et al [54].

The biological implications of the 3-grams distinguished by high scarcity scores were examined using the dataset of 59 enzyme/ligand complexes compiled by Gutteridge and Thornton [24] (GT dataset; see Table 1). We performed three case studies: C-Src kinase, hemoglobin subunits α and β, and tyrosyl-tRNA synthetase. Homologous sequences were retrieved in each case by a PBLAST search [55] of the SwissProt database. Sequences with E-values lower than 10^{-57} were extracted, leading to m = 111 homologous sequences for c-Src, 138 and 224 sequences for hemoglobin α- and β-subunits, respectively, and 24 sequences for tyrosyl tRNA synthetase. The sequences were aligned with ClustalX using default parameters [56].

**Exposed 3-grams**

The Accessible surface area (ASA) of each amino acid in the GT database was calculated using the program NACCESS [57]. Amino acids were defined as exposed if their relative exposed surface area was > 20%. 3-grams with two or more exposed amino acids were defined as exposed ones.

**Authors’ contributions**

DT did the calculations. DT and IB analyzed the results and wrote the manuscript. Both authors read and approved the final manuscript.

**Additional material**

Additional File 1
Supplementary tables S1-S7. Figures and material. Click here for file [http://www.biomedcentral.com/content/supplementary/1471-2105-8-226-S1.doc]

**Acknowledgements**

We thank Thomas E. Smithgall for helpful discussions on the structure-function of Src proteins family. Support by NIH R01 grant # LM007994 is gratefully acknowledged.

**References**

1. Casari G, Sander C, Valencia A: A method to predict functional residues in proteins. Nat Struct Biol 1995, 2:171-178.
2. Lichtarge O, Bourne HR, Cohen FE: An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996, 257:342-358.
3. Hannenhalli SS, Russell RB: Analysis and prediction of functional sub-types from protein sequence alignments. J Mol Biol 2000, 303:61-76.
4. Armon A, Graur D, Ben-Tal N: ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 2001, 307:447-463.
5. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N: Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 2002, 18 Suppl 1:S71-S77.
6. Andrade MA, Casari G, Sander C, Valencia A: Classification of protein families and detection of the determinant residues with an improved self-organizing map. Biol Cybern 1997, 76:441-450.
7. Aloy P, Querol E, Aviles FX, Sternberg MJ: Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol 2001, 311:395-408.
8. Madabushi S, Yao H, Marsh M, Kristensen DM, Filipppi A, Sowa ME, Lichtarge O: Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J Mol Biol 2002, 316:139-154.
9. Guttridge A, Bartlett GJ, Thornton JM: Using a neural network and spatial clustering to predict the location of active sites in enzymes. J Mol Biol 2003, 330:179-194.
10. Panchenko AR, Kondrashov F, Bryant S: Prediction of functional sites by analysis of sequence and structure conservation. Protein Sci 2004, 13:884-892.
11. Landgraf R, Xenarios I, Eisenberg D: Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 2001, 307:1487-1502.
12. Wallace AC, Borkakoti N, Thornton JM: TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci 1997, 6:2308-2323.
13. Fetrow JS, Godzik A, Skolnick J: Functional analysis of the Escherichia coli genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. J Mol Biol 2001, 313:283-295.
14. Kleywegt GJ: Recognition of spatial motifs in protein structures. J Mol Biol 1999, 285:1887-1897.
15. Shen SC, Bahar I: Mining frequent patterns in protein structures: a study of protease families. Bioinformatics 2004, 20 Suppl 1:S17-S25.
16. Yang LW, Bahar I: Coupling between catalytic site and collective dynamics: a requirement for mechanochanical activity of enzymes. Structure 2005, 13:893-904.
17. del Sol A, Fujishita H, Amoros D, Nussinov R: Residues crucial for maintaining short paths in network communication mediate signaling in proteins. Mol Syst Biol 2006, 2:0006.
18. Chennubhotla C, Bahar I: Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES. Mol Syst Biol 2006, 2:0136.
19. Vries JK, Munchi T, Tobö D, Klein-Seetharan J, Benos PV, Bahar I: A sequence alignment-independent method for protein classification. Appl Bioinformatics 2004, 3:137-148.
20. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magramane M, Martin MJ, Ntallie DA, O'Donovan C, Redaschi N, Yeh LS: The Universal Protein Resource (UniProt). Nucleic Acids Res 2005, 33:D154-D159.
21. Herzberg O, Moult J: Analysis of the steric strain in the polypeptide backbone of protein molecules. Proteins 1991, 14:223-229.
22. Petock JM, Torshin YI, Weber IT, Harrison RW: Analysis of protein structure reveals regions of rare backbone conformation at functional sites. Proteins 2003, 53:872-879.
23. Novotny M, Kleywegt GJ: A survey of left-handed helices in protein structures. J Mol Biol 2005, 347:121-241.
24. Guttridge A, Thornton JM: Conformational changes observed in enzyme crystal structures upon substrate binding. J Mol Biol 2005, 346:21-28.
25. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM: Analysis of catalytic residues in enzyme active sites. J Mol Biol 2002, 324:105-121.
26. Brown MT, Cooper JA: Regulation, substrates and functions of src. Biochem Biophys Res Commun 2004, 324:1155-1164.
27. Lerner EC, Smithgall TE: SH3-dependent stimulation of Src-family kinase autophosphorylation without tail release from the SH2 domain in vivo. Nat Struct Biol 2002, 9:365-369.
28. Feng S, Kasahara C, Ricles R, Schreiber SL: Specific interactions outside the proline-rich core of two classes of Src homology 3 ligands. Proc Natl Acad Sci U S A 1995, 92:12048-12051.
29. Okihio N, Tanaka T, Fukuda R, Nagai M: Differential ligand recognition by the Src and phosphatidylinositol 3-kinase Src homology 3 domains: circular dichroism and ultraviolet resonance Raman studies. Biochemistry 2003, 42:208-216.
30. Young MA, Gonfioli S, Spectri-Furga G, Roux B, Kuriyan J: Dynamic coupling between the SH2 and SH3 domains of c-Src and Hck underlies their inactivation by C-terminal tyrosine phosphorylation. Cell 2001, 105:115-126.
31. Oshkirim E, Post CB: Src kinase activation: A switched electrostatic network. Protein Sci 2006, 15:1051-1062.
32. Hulo N, Bairoch A, Bailloud V, Cerutti L, de CE, Langendijk-Jenaeus PS, Pagini M, Sigrist CJ: The PROSITE database. Nucleic Acids Res 2006, 34:D227-D230.
33. Changeux JP, Edelstein SJ: Allosteric mechanisms of signal transduction. Science 2003, 308:1424-1428.
34. Kleywegt GJ, Johnson ML, Adams GK: Oxygentiation-subunit interactions in human hemoglobin: experimental studies on the concentration dependence of oxygenation curves. Biochemistry 1976, 15:5330-5362.
35. Mariniucci M, Mavilio F, Massa A, Gabbianelli M, Fontanarosa PP, Camagna A, Iveciti C, Tontoni L: A new abnormal human hemoglo- bin: Hb Prato (alpha 2.3 (B12) Arg leads to Ser beta 2). Biochim Biophys Acta 1979, 578:534-540.
36. Kavanaugh JS, Rodgers PH, Arnone A, Hui HL, Wierzb A, DeYoung A, Kwiatkowski LD, Noble RW, Juszczak LJ, Peterson ES, Friedman JM: Intersubunit interactions associated with Tyr4a alpha sta- bilize the quaternary-T tetramer but are not major quater- nary constraints in deoxyhemoglobin. Biochemistry 2005, 44:3806-3820.
37. Barrick D, Ho NT, Simpleanu V, Dahlquist FW, Ho C: A test of the role of the proline-rich core in the Perautz model for cooperation in haemoglobin. Nat Struct Biol 1997, 4:78-83.
38. Bonaventura C, Bonaventura J: Anionic control of hemoglobin function. In Biochemical and Clinical Aspects of Hemoglobin Abnormali- ties Edited by: Caughey WS, New York: Academic; 1978:647-663.
39. Perutz MF, Wilkinson AJ, Paoli M, Condó P, Kleywegt GJ: Thermodynamic studies on the equilib- rium properties of a series of recombinant betaW37 hemoglo- bin mutants. Biochemistry 1998, 37:4336-4345.
40. Elion J, Wajeman H, Belkhodja-Dunda O, Labie D, Meserschmidt J, Saal AM, Desabiers B: [Hemoglobin J: Anmiens betabeta17 (A 14) Lys was replaced by Asn. Linking properties of a func- tionally silent new abnormal hemoglobin and a polyclonemia vera (author's trans)]. Nouv Rev Fr Hematol 1979, 21:347-352.
41. Kiger L, Klinger AL, Kwiatkowski LD, De YA, Doyle ML, Holm JT, Noble RW, Ackers GK: Thermodynamic studies on the equilib- rium properties of a series of recombinant betaW37 hemoglo- bin mutants. Biochemistry 1999, 38:7253-7243.
42. Kavanaugh JS, Weydart JA, Rodgers PH, Arnone A, Hui HL, Wierzb A, Kwiatkowski LD, Pauly P, Noble RW, Bruno S, Mozarelli A: Site-directed mutations of human hemoglobin at residue 35beta: a residue at the intersection of the alpha1beta1, alpha1beta2, and alpha1alpha2 interfaces. Protein Sci 2001, 10:1847-1857.
43. Richard V, Dodson GG, Mauguen Y: Human deoxyhaemoglobin 2-3-diphosphoglycerate salt structure at 2.5 A resolution. J Mol Biol 1993, 233:270-274.
44. Aime S, Digilio G, Fassano M, Paletti S, Arnelli A, Ascenzi P: Metal complexes as allosteric effectors of human hemoglobin: an NMR study of the interaction of the gadolinium(III) bis(m-benzoxyphenyl amino)ethane salt with HbWA73 hemoglobin. Biochemistry 1999, 38:7623-7634.
45. Xu C, Tobö D, Bahar I: Allosteric changes in protein structure computed by a simple mechanical model: hemoglobin T---> R transition. J Mol Biol 2003, 333:153-166.
46. Baldwin J, Chothia C: Haemoglobin: the structural changes related to ligand binding and its allosteric mechanism. J Mol Biol 1979, 129:175-220.
47. Brick P, Bhat TN, Blott DM: Structure of tyrosyl-tRNA synthetase refined at 2.3 A resolution. Interaction of the enzyme with the tyrosyl adenylate intermediate. J Mol Biol 1989, 208:83-98.
48. Fersht AR, Shi JP, Wilkinson AJ, Blow DM, Carter P, Way MMY, Winter GP: Analysis of Enzyme Structure and activity by protein engineering. Angewandte Chemie 1994, 34:2223-2228.
49. First EA, Fersht AR: Analysis of the role of the KMSKS loop in the catalytic mechanism of the tyrosyl-tRNA synthetase using mutagenic cycles. Biochemistry 1995, 34:5030-5043.
50. Shannon CE: The mathematical theory of communication. The Bell system technical Journal 1948, 27:379-423 and 623-656.
51. Venter JC, Adams MD, Myers EW, Li PW, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng HO, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor, D227-D230.
M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di F V, Dunn P, Elbeck K, Evangelista C, Gabrielson AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lei Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Liu F, Merkulov GV, Milhina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Daniel S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Fernis S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannke C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abruil JF, Guigo R, Campbell M, Sjolander KV, Karlak B, Kejarwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diener K, Murugunujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yoossph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Cauikl P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Foster C, Gire H, Glanowski S, Glasser K, Glokeck C, Grompkhov M, Graham K, Grapman D, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha K, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, M A, Majors W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M: The sequence of the human genome. Science 2001, 291:1304-1351.

53. Brenner SE: A tour of structural genomics. Nat Rev Genet 2001, 2:801-809.

54. Ganapathiraju G, Weisser D, Klein-Seetharaman J, Rosenfeld R, Carbonell J, Reddy R: Comparative n-gram analysis of whole-genome sequences. San Diego. 2002.

55. Altschul SF, Madden TL, Schaffer AA, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997, 25:3389-3402.

56. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997, 25:4676-4682.

57. Hubbard SJ, Thornton JM: NACCESS, Computer Program. Department of Biochemistry and Molecular Biology, University College, London 1993.

58. Xu W, Doshi A, Lei M, Eck MJ, Harrison SC: Crystal structures of c-Src reveal features of its autoinhibitory mechanism. Mol Cell 1999, 3:629-638.

59. DeLano WL: DeLano, W.L. The PyMOL Molecular Graphics System (2002) DeLano Scientific, Palo Alto, CA, USA. 2005.

60. Tame JR, Vallone B: The structures of deoxy human haemoglobin and the mutant Hb Tyralpha42His at 120 K. Acta Crystallogr D Biol Crystallogr 2000, 56:805-811.