Subgroup-Rich and Subgroup-Poor Substitution Groups in Protein Motifs

Technical Report, Dept of CSE, Wright State University

Guozhu Dong, Vineela Muppavarapu, and Gauri Dandekar

Department of Computer Science and Engineering
Wright State University
Dayton, OH 45435, U.S.A.
gdong@cs.wright.edu

August 27, 2004

Abstract

Motivation: Protein motifs characterize protein domains and are important for determining the structure and function of proteins from sequences. Substitution groups are sets of amino acids that can substitute each other in similar contexts, and they are important components of the motifs.

Results: This paper provides certain extreme substitution groups, namely the subgroup-rich and the subgroup-poor ones, derived from the protein motifs from the PROSITE database; these groups have unusually high or low numbers of subsets that also appear as substitution groups in the motifs. It also provides some statistics on the substitution groups. These results may provide insights on questions such as how flexible and how tightly bonded substitution groups are, and how amino acids interact; they may even suggest new types of substitution groups and motifs that still need to be discovered.

Keywords: amino acid substitution groups, protein motifs subgroup-rich/poor, substitution flexibility, substitution group statistics, subgroup occurrence distribution.

*Part of this work was supported by the Genome Research Infrastructure Partnership Project from BRTT, Ohio. Correspondence should be addressed to Guozhu Dong.
1 Background

Protein motifs are specific regions of protein sequences that are conserved in both structure and sequence. These regions are generally important for the function of a protein and for the maintenance of its three-dimensional structure. Motifs are useful for identifying the protein family, and the structure and function, of unknown protein sequences.

A major type of component in motifs is the substitution group. A substitution group is a set of amino acids that share certain common properties and can substitute each other in particular positions and contexts in protein sequences. For example, $[R K] (2) - x - [S T]$ is a motif. It contains two substitution groups, namely $[R K]$ and $[S T]$. The square brackets show the beginning and end of the substitution groups. The motif spans over 4-positions: the first position can have either R or K, the second either R or K, the third is arbitrary, and the last position either S or T. The number 2 indicates that the first substitution group occurs twice.

In this paper we use the motifs and substitution groups given in Release 17.01 of the PROSITE database available at www.expasy.org/prosite.

Related works in bioinformatics can be divided into studies on motifs (e.g. [3, 2, 4]), studies on the discovery of substitution groups (e.g. [6]), and studies using properties of substitution groups (e.g. [5]). We note that [4] uses statistical over- or under-representation to determine whether a motif is important; it does not provide statistics on the motifs.

2 General Statistics of the Substitution Groups

The statistics are based on the length$^1$ of substitution groups. The average length of all distinct substitution groups is 4. The count distribution of distinct substitution groups over length – the number of distinct groups for given lengths – is given in Figure 1. Note that length 4 has the largest count of 285.

The frequency distribution of the most frequent substitution group for given lengths is given in Figure 2. Table 1 shows the most frequent substitution groups for different lengths. For all lengths, the most frequent substitution group is [ILVM] and it occurs 752 times. The are 9 most frequent substitution groups for length 11, all with the frequency of 1: [ACFGILMSTVY], [AFGHKLPQSVY], [AFGILMSTWVY], [ACFGILMSTVV], [ACFILMSTVYW], [ACFGILMNSTV], [ACFGKNQRSTVY], [DFHILNRSTVY], [DFHILNRSTVY], [DFHILNRSTVY], [DFHILNRSTVY], [DFHILNRSTVY], [DFHILNRSTVY]...

$^1$The terms cardinality and length are used interchangeably to refer to the number of amino acids in a substitution group.
[FGILMPQSVWY].

There are three interesting observations: (1) From Table 1 we see that the most frequent substitution groups for lengths between 3 and 8 are “monotone”, in the sense that the most frequent substitution group for length $i + 1$ is the result of adding an additional amino acid to that of length $i$; the additional amino acids for lengths from 4 to 8 respectively are $M, F, Y, W, A$. (2) Also from Table 1 we see that all of the most frequent substitution groups for lengths between 4 and 10 contain [ILVM]. (3) The two distributions in Figures 1 and 2 both peak at 4, which is also the average length of the substitution groups.

3 Subgroup-Rich/Poor Substitution Groups

We now characterize substitution groups based on the count distribution of their subsets which also occur as substitution groups in the PROSITE mo-
<table>
<thead>
<tr>
<th>Length</th>
<th>Most Frequent</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>[KR]</td>
<td>228</td>
</tr>
<tr>
<td>3</td>
<td>[ILV]</td>
<td>188</td>
</tr>
<tr>
<td>4</td>
<td>[ILMV]</td>
<td>752</td>
</tr>
<tr>
<td>5</td>
<td>[FILMV]</td>
<td>261</td>
</tr>
<tr>
<td>6</td>
<td>[FILMVY]</td>
<td>157</td>
</tr>
<tr>
<td>7</td>
<td>[FILMVWY]</td>
<td>61</td>
</tr>
<tr>
<td>8</td>
<td>[AFILMVWY]</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>[ACGILMSTV]</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>[ACFILMSTVY]</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>[ACFGILMSTVW]</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>[ACFGILMSTVYW]</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>[ACFGHILMPQTVYW]</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1: Frequency of Most Frequent Substitution Groups over Length

tifs. More specifically, we determine the average distribution for substitution groups for each length, and we consider substitution groups with unusually high/low distributions.

We believe that these results can be used to characterize how flexible substitution groups are, or how tightly the amino acids are bonded, in these groups. We elaborate on these points below.

Let $\mathcal{G}$ be the collection of all substitution groups that occur in the PROSITE database. For each group $x \in \mathcal{G}$ and each positive integer $2 < i < |x|$, let $s(x, i) = \frac{|\{y \in \mathcal{G} \mid x \subset y, |y| = i\}|}{C_i^{|x|}}$. That is, $s(x, i)$ is the fraction of number of subsets $y$ of $x$ of length $i$ such that $y$ also occurs as a substitution group, over the number of possible subsets of $x$ of length $i$. We call $(s(x, 2), s(x, 3), \ldots, s(x, |x| - 1))$ the subgroup occurrence distribution of $x$.

We use a small example to illustrate. Table 2 shows the distribution of the substitution group [EILMV]. The group has length 5. For each length between 2 and 4, we list all subsets of the group of that length; "*" indicates that the corresponding subset occurs as a substitution group. As can be seen, the subgroup occurrence distribution for [EILMV] is $(0.60, 0.50, 0.40)$.

The typical subgroup occurrence distribution for each length is defined as the average subgroup occurrence distribution over all substitution groups of that length. We denote the typical subgroup occurrence distribution for a

---

$^2|x|$ denotes the cardinality of $x$. 

4
<table>
<thead>
<tr>
<th>Length</th>
<th>Subgroups</th>
<th>( s([EILMV], \text{Length}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>[EI] [EL] [EM] [EV] [IL]*</td>
<td>0.60 = ( \frac{6}{10} )</td>
</tr>
<tr>
<td>3</td>
<td>[IM]* [IV]* [LM]* [LV]* [MV]*</td>
<td>0.50 = ( \frac{5}{10} )</td>
</tr>
<tr>
<td>4</td>
<td>[EILM] [EILV] [EIMV] [ELMV]*</td>
<td>0.40 = ( \frac{4}{10} )</td>
</tr>
</tbody>
</table>

Table 2: Illustration: Subgroup Occurrence Distribution for [EILMV]

length \( \ell \) using a function \( \tau(\ell, i) \) (for \( 2 < i < \ell \)) as

\[
\tau(\ell, i) = \sum_{x} \frac{s(x, i)}{n},
\]

where \( x \) ranges over substitution groups in \( G \) of length \( \ell \) and \( n \) is the number of such \( x \).

Figure 3 plots the typical subgroup occurrence distributions for different group lengths. The figure shows that small subsets of substitution groups frequently occur as natural substitution groups, but their large counterparts occur as substitution groups much less frequently.

![Figure 3: Typical Subgroup Occurrence Distributions](image)

We measure extreme subgroup occurrence distributions using two numerical measures. The first, called the *subgroup density measure* and denoted as \( \delta_{sg} \), is defined as

\[
\delta_{sg}(x) = \sum_{i=2}^{\frac{|x|-1}{2}} s(x, i)
\]
for each group \( x \). The second, called the subgroup distribution deviation measure and denoted as \( \text{devAvg} \), is defined as

\[
\text{devAvg}(x) = \sum_{i=2}^{\lfloor |x| - 1 \rfloor} |s(x, i) - \tau(|x|, i)|.
\]

For each given length, a group with the highest subgroup density is called a subgroup-rich group. Similarly, a group with the lowest subgroup density is called a subgroup-poor group. A group \( x \) with highest \( \text{devAvg}(x) \) among all group of length \( |x| \) is called a deviate-most group.

Table 3 contains, for each length, the typical subgroup occurrence distribution (the T row), the subgroup-rich groups (SR), the subgroup-poor groups (SP), and the deviate-most groups (DM). We use * to indicate that the SP or SR group is also the DM group. The first column is the length. The numbers in each subgroup occurrence distribution are listed in increasing length order from 2 to \( |x| - 1 \) (where \( |x| \) is the length of the group under consideration). The table does not include entries for the largest length of substitution groups, namely 13, since there is only one substitution group for this length. We discuss the situation for lengths 3 and 4 in Section 3.1.

The table clearly shows that the subgroup-rich groups have larger frequency values in their subgroup occurrence distributions than the subgroup-poor groups, and the subgroup-poor groups have more values close to 0's. A deviate-most group is almost always either a subgroup-rich group or a subgroup poor group, except for length 11.

We now elaborate on how the extreme distributions are conjectured to relate to flexibility and tight bondedness of substitution groups, and how one may use such distributions to guide search for other protein domains and proteins. First, a subgroup-rich group can be viewed as a very flexible substitution group, since by definition it has unusually large number of subsets which can occur as substitution groups by themselves. Second, a subgroup-poor group can be viewed as a very tightly bonded substitution group, since by definition it has much fewer subsets which occur as substitution groups than normal and in many situations the amino acids of the subsets cannot substitute each other. Third, a subgroup-poor group can be used to suggest possibilities of proteins or motifs which have not been discovered yet – perhaps the fact of relatively smaller number of subsets which occur as substitution groups was caused by our current lack of knowledge and not by nature.

We also considered a different way to define the density measure by taking into account the fact indicated by Figure 1 – there are more distinct substitution groups of certain lengths than others. We did this by normalizing
s(x, i), that is replacing s(x, i) with $\frac{s(x, i)}{\text{count}(i)}$ where count(i) is the number of distinct groups of length i. This only led to minor changes to Table 3 and the details are omitted.

### 3.1 Subgroup-Rich/Poor for Lengths 3 and 4

For lengths 3 and 4, there are many subgroup-rich groups. For length 4, there are 38 subgroup-rich groups, all having the subgroup occurrence distribution of (1.00, 1.00): [ACGS], [ACLV], [ACST], [ACSV], [ADEC], [ADGS], [ADST], [AESG], [AGNS], [AGPS], [AGQS], [AGST], [AGSV], [AILV], [AITV], [ALMV], [ALPS], [ANQS], [ANST], [APST], [ASTV], [DEGS], [DEST], [DGNS],
[DKNR], [DNST], [EGST], [EKQR], [FILV], [FILY], [FLMY], [FLVY], [GHNS], [ILMV], [KNQR], [KQRS], [KRST], [NQRS]. There is just one subgroup-poor group, namely [DSWY], with the distribution of (0.50, 0.00), which is also the deviate-most group. For length 3, there are 185 subgroup-rich groups, all having the subgroup occurrence distribution of (1.00): [ACG], [ACL], [ACP], [ACS], [ACT], [ACV], [ADE], [ADG], [ADP], [ADS], [ADT], [ADV], [AEG], [AEH], [AEP], [AFG], [AFP], [AFS], [AFV], [AFY], [AGM], [AGN], [AGP], [AGQ], [AGR], [AGS], [AGT], [AGV], [AHK], [AHM], [AHN], [AIL], [AIT], [AIV], [AKP], [AKR], [ALM], [ALP], [ALS], [ALV], [AMV], [ANQ], [ANS], [ANT], [APQ], [APS], [APT], [APV], [AQS], [ART], [AST], [ASV], [ATV], [AWY], [CDG], [CDN], [CDS], [CFY], [CGN], [CGS], [CHS], [CLV], [CST], [CSV], [DEG], [DEH], [DEK], [DEN], [DEQ], [DER], [DES], [DET], [DGH], [DGN], [DGQ], [DGS], [DGT], [DHN], [DKN], [DKR], [DNQ], [DNR], [DNS], [DNT], [DNV], [DST], [DSV], [EGK], [EGP], [EGQ], [EGR], [EGS], [EGT], [EHQ], [EHS], [EKP], [EKQ], [EKR], [EPR], [EPS], [EQR], [EQS], [EST], [FHR], [FHY], [FIL], [FIV], [FIY], [FLM], [FLT], [FLV], [FLY], [FMY], [FTY], [FVW], [FVY], [FWY], [GHN], [GHQ], [GHS], [GIV], [GKN], [GKR], [GKS], [GLN], [GLV], [GNQ], [GNS], [GPS], [GQR], [GQS], [GRS], [GRT], [GST], [GSV], [HKR], [HQN], [HNS], [HNY], [HPQ], [HQR], [HQT], [HQY], [HRY], [HST], [HWY], [ILM], [ILT], [ILV], [ILY], [IMV], [ITV], [IWV], [KNQ], [KNR], [KPO], [KPR], [KQR], [KQS], [KRS], [KRT], [KRW], [KST], [LMV], [LMY], [LNQ], [LNS], [LPS], [LPV], [LQS], [LST], [LVY], [NQR], [NQS], [NRS], [NST], [PQR], [PQT], [PST], [PSV], [QRS], [QRT], [QST], [RST], [STV]. There are 9 subgroup-poor groups, all with the subgroup occurrence distribution of (0.33): [CIN], [CKR], [DIN], [EFY], [EIV], [GSW], [KMR], [KRV], [MST]. These are also the deviate-most groups for length 3.

Website for Additional Information

The following additional information is available at the following website http://www.cs.wright.edu/~gdong/motif: (1) The original PROSITE data, the extracted motifs, and the extracted substitution groups; (2) the count of substitution groups for given lengths and the frequency for substitution groups; and (3) subgroup occurrence distribution for the substitution groups.

References


