Automated Symmetry and Quasisymetry Detection in Protein Structure

Anna K. Labno*
MIT Department of Biology

(Dated: May 17, 2007)

Symmetry is a persistent and understudied theme in biology, especially visible in proteins' three dimensional structure. In this paper we present mathematical groundwork underlying symmetry detection and apply it to obtain computerized symmetry detection scheme allowing for detection of symmetries and quasi-symmetries in three dimensional protein structures with desired degree of exactness. The algorithm allows for extraction of translation, rotation, reflection and uniform scaling and extracts exact and partial symmetries at all scales as shown by testing it on sample of protein structures taken from PDB. Such routines can facilitate systematic studies of structural symmetries and help elucidate their role in biology.

1. INTRODUCTION

Structural symmetry of protein three dimensional structure has long been a fascinating phenomenon. The majority of both soluble and membrane-bound proteins are symmetrical, which is especially true for oligomeric complexes. Those proteins adopt only enantiomorphic symmetries [1] since mirror symmetry or inversion, if present, would necessitate presence of D-amino acids which are not found in nature. Symmetrical proteins are thought to have increased stability of association as it is thought that the symmetrical structure generally is the lowest energy state one [2, 3], allow for formation of oligomers of defined copy number which prevents from unwanted aggregation [1] and are hypothesized to have fewer kinetic barriers to folding than symmetric structures [4]. Some symmetrical structures are thought to use their symmetry in order to perform certain morphological function, for example transcription factors use symmetry to enhance specificity of allosteric regulation they extert [5]. The role of quasisymmetry and pseudosymmetry, often present in viral capsids is not clearly understood but it is generally thought to be caused by need of greater structural support [1]. It was also suggested that evolution favors symmetrical structures [6] however there have been a couple of examples where evolution, perhaps driven by functional need, moved from symmetrical to asymmetrical structure, such as in case of photosystem I [7]. Analysis of symmetry of proteins has also served as a basis for studying emergence of novel protein topologies [8, 9] with special attention to the situation where functional constrains compete with maintenance of global symmetry [10] . Symmetry has also been the basis for most acknowledged allostery model in which switching between two or more symmetrical states of molecule occurs in a response to stimuli.

Surprisingly, for such prevalent topic, no systematic study utilizing wealth of protein data deposited in PDB was conducted to the best of our knowledge. There has been a systematic study of major classes of protein structures as defined by their secondary structure (α/β; α+β, all β and all α) which shed some light on the evolution of protein fold [11]. Similar analysis based on symmetries exhibited by protein three dimensional structure could yield more valuable insights into evolution and structure-function relationship in proteins. However in order for such an assessment one needs an automated, computerized symmetry detection tools, which are capable of large scale protein testing in relatively short time. In this paper we have adopted existing symmetry detection algorithm [12] which is based on matching simple local shape signatures which are then clustered in a way that allows extracting symmetrical features. This code has been tested on both monomeric and oligomeric structures taken from PDB and shown to allow for fast detection of exact and approximate symmetries.

2. THEORY OF SYMMETRY DETECTION

While in our eyes symmetries are an obvious element of Nature, they are difficult to handle mathematically because of the fact that they need not be simple symmetries such as reflections or rotations, but fairly often we deal with composite ones—parts of objects are both rotated and reflected, or rotated and translated. Complexity of the problem rises even further due to the imperfections of Nature: symmetries, even though clearly visible are imperfect and thus generally do not have an elegant mathematical description. However, in the following mathematical description of Euclidean symmetries we will assume existence of only exact symmetries and will relax this assumption in the algorithm discussion.

Consider two points P and Q with $\vec{x}$ and $\vec{y}$ in a common Cartesian coordinate system $\mathcal{C} = (\hat{\imath}, \hat{\jmath}, \hat{k})$, each with a distinguished set of orthonormal axes (local coordinate system) $\mathcal{C}[P]$ and $\mathcal{C}[Q]$, respectively. We say that the points P and Q are symmetric with respect to transformation $\mathcal{T}$ if

$$\mathcal{T}\left[\mathcal{C}[P]\right] \longrightarrow \mathcal{C}[Q], \qquad (1)$$

i.e. if the transformation $\mathcal{T}$ transforms the set of orthonormal axes of P onto the set of orthonormal axes of

*Electronic address: labnoa@mit.edu

Q. While there is a large number of possible transforms—they could be composite transforms—it turns out that all possible transforms can be generated by three primary transforms: translation, rotation, and uniform scaling. Thus, given two sets of orthonormal axes one can easily calculate the primitive transforms and recover the composite transform $\mathcal{T}$ due to the fact that rotation and uniform scaling are linearly independent from translation.

Therefore, the total transform $\mathcal{T}$ can be represented as a vector in seven-dimensional vector space

$$\mathcal{T} = (s, R_x, R_y, R_z, t_x, t_y, t_z), \quad (2)$$

where s is the uniform scaling parameter (negative for reflection), (R_x, R_y, R_z) are the Euler angles describing rotation, and the vector $\vec{t} = (t_x, t_y, t_z)$ describes translation. The transformation, when applied to a vector $\vec{v}$, transforms it into

$$\vec{v}' = \vec{t} + sR\vec{v}, \quad (3)$$

with R being the rotation matrix constructed from the three Euler angles. Therefore, we need a way of estimating the three parameters, namely the translation, rotation, and uniform scaling. Because of the fact that we deal only with the carbon chain approximated by a series of points, we may easily obtain the translation as

$$\vec{t} = \vec{y} - \vec{x}. \quad (4)$$

Similarly, the rotation can be extracted purely from the consideration of relative orientations of the two sets of orthonormal axes related to P and Q. As it turns out, we can simply use the fact that R would be a matrix related to change of basis, and hence

$$R = F_q F_p^{-1} = \begin{pmatrix} \uparrow & \uparrow & \uparrow \\ \hat{n}_q & \hat{p}_q & \hat{c}_q \\ \downarrow & \downarrow & \downarrow \end{pmatrix} \begin{pmatrix} \uparrow & \uparrow & \uparrow \\ \hat{n}_p & \hat{p}_p & \hat{c}_p \\ \downarrow & \downarrow & \downarrow \end{pmatrix}^{-1}, \quad (5)$$

where $\hat{n}$ is the normal vector, $\hat{p}$ is the perpendicular vector, and $\hat{c}$ is the co-perpendicular vector for a given amino acid. The three are defined in a simple way. The normal vector is the one parallel to the chain, i.e. pointing along the chain at the position of the amino acid. The perpendicular vector is a vector perpendicular to the place created by the bonds made by the current amino acid with its two direct neighbors. Finally, the co-perpendicular vector $\hat{c}$ is just a vector product of the previous ones, namely

$$\hat{c} = \hat{n} \times \hat{p}. \quad (6)$$

However, we have not yet mentioned the reflection symmetry specified by uniform scaling parameter s. Due to the fact that there are physical bounds on sizes of protein structures, namely on the sizes of alpha helices and beta sheets, the value of s has to be limited to either 1 or -1. We also realize, that if $s = -1$ it will be impossible to extract meaningful Euler angles from the matrix R. Therefore, we could simply assume $s = 1$ and calculate the Euler angles from R with a subsequent verification step. In case of failed verification, this would imply that $s = -1$, thus specifying all the parameters of the transformation.

3. ALGORITHM FOR AUTOMATED SYMMETRY FINDING

The complete algorithm follows loosely the approach of [12], who in turn was inspired by previous work of Hough on voting algorithms for detection of shapes in the two-dimensional case [13]. Given that all mathematical background has been already derived in the preceding section, we will discuss only the algorithmic frame here.

We begin with calculation of the axes specifying the local coordinate system and thus orientation of each amino acid (see Alg. 1). Those are then used to obtain a set of transformations $\mathcal{T}_{ij}$ for each pair of amino acids in the protein chains being compared. Thus, for each pair we find a symmetry that could be responsible for the two amino acids to be geometrically connected, and each transformation found this way acts as a vote: the more votes are given on similar transformation, the more likely it is that a symmetry corresponding to the given transformation indeed exists in the system.

In order to obtain the dominating symmetries and disregard the errors coming from approximation, we use the mean-shift algorithm to find the modes of the transformation distribution $\mathcal{T}_{ij}$. Those are then counted as meaningful symmetries and used to determine patches of the protein chains connected through the given symmetry, which finishes the algorithm.

4. TESTING SAMPLE

In order to test the algorithm we chose a couple of representative but relatively small proteins from Protein Data Base (PDB) due to constrains in computational power. Attempt was also made to choose proteins in which symmetry, or its lack, can be clearly seen by eye in order to verify the results of the computations.

In N-terminal fragment of NS1 protein from Influenza virus (1AIL), which is a small protein encoded by the virus which prevents activation of NF-κB and induction of α/β interferon [14], (Figure 1a) we found that the the molecule is internally symmetrical with its three α-helices split evenly between the two symmetrical subunits. The symmetry was primary a reflection with respect to a central plane followed by slight rotation (1.7 rad around horizontal x-axis) and translation of the second (green subunit) by -4.5nm in x direction and -1.1nm in y direction with respect to the first (blue) subunit. The appearance of mirror symmetry although initially surprising can be explained by the fact that right-handed and left-handed helices differ only slightly in the absolute spa-

Algorithm 1 Symmetry finding between two proteins.

```
Require: Two chains of amino acids, $a_i$ and $b_i$, and amino acids positions in Cartesian coordinates, $\vec{x}_i$ and $\vec{y}_i$.
Ensure: The set of symmetries realized through unknown transformations $\mathcal{T}_{ij}$ and the fragments of protein sequences interrelated through symmetries.
for each chain do
    Calculate the "normal" vectors $\vec{n}_i$ pointing along the chain for each element in the chain.
    Calculate the perpendicular vector $\vec{p}_i$ normal to the plane formed by the two neighboring residues.
    Calculate the co-perpendicular vector $\vec{c}_i = \vec{n}_i \times \vec{p}_i$.
    Compute a transformation matrix $F_i = [\vec{n}_i \vec{p}_i \vec{c}_i]$.
end for
for $i := 0$ to $n$ do
    for $j := 0$ to $n$ do
        $\vec{t}_{ij} := \vec{y}_j - \vec{x}_i$
        $R := F_j F_i^{-1}$
        if $R$ does not realize a pure rotation then
            $s := -1$
        else
            $s := 1$
        end if
        Compute the Euler angles $(R_x, R_y, R_z)$ from the matrix $R$
        $\mathcal{T}_{ij} = [s, R, \vec{t}]$
    end for
end for
Calculate modes of the distribution of $\mathcal{T}_{ij}$ using mean-shift algorithm in order to obtain the most frequent transformations present in the protein.
for each frequent transformation do
    Find the pair of amino acids $ij$ connected by a transformation most closely resembling the given modal transformation.
    Use patch growing algorithm around the found pair in order to find the symmetric fragments of a given protein with respect to the specified modal transformation $\mathcal{T}_{ij}$
end for
```

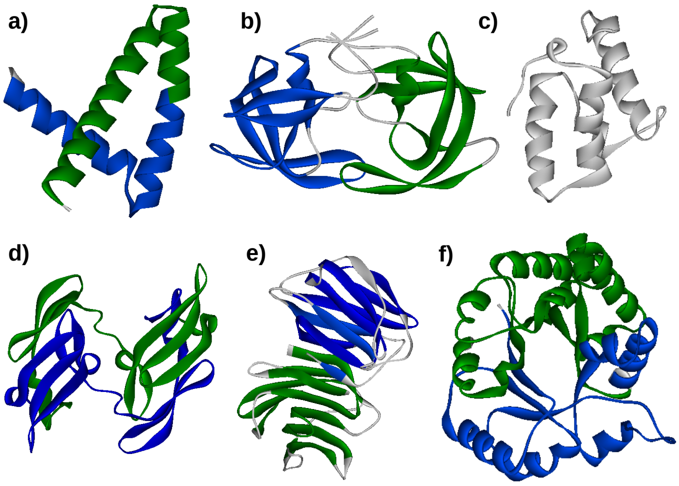

FIG. 1: Symmetries detected by the computerized detection scheme in A. NS1 protein from Influenza virus (1AIL) B. HIV-1 protease (1EBY) C. colicin E8 inhibitor D. Cyanovirin-N (1J4V) E. S-Lectin (1SLT) F. triosephosphate isomerase (1IF2)

tial arrangement of molecules of carbon chain. In order to find large-scale symmetries the voting algorithm averages over votes in way that allows for non-exact symmetries to be determined, which results in detection of two L-amino acid α helices as mirror symmetries of each other. If the bandwidth is small only the symmetries within secondary structure, translations and rotations will be found (data not shown), but the overall quasi-symmetries in proteins are extremely common and could have functional and evolutionary meaning, hence the bandwidth was kept sufficiently large for those symmetries to be observed. Subsequently the algorithm was tested on proteins with more complex structure, such as dimeric HIV-1 protease (1EBY), which cleaves the nascent polyproteins during viral replication [15]. In this protein the two monomers are related to each other by reflection followed by a rotation around y-axis by -10rad and z-axis (out of the page) by -18rad. Small, α-helical asymmetric molecule, colicin E8 inhibitor (1GXG) which is postulated to inhibit activity of colicin E8 speculated to affect gene structure as a part of immune response [16]. No symmetry was detected in this protein as shown on Figure 1c. Two other dimeric protein, consisting mostly β-sheets but with significant content of random loops were tested: Cyanovirin-N (1J4V) (human immunodeficiency virus-inactivating protein [17]) and S-Lectin (1SLT) (β-Galactoside-Binding Protein citeLiao) as shown on Figure 1d and e. It is interesting to notice that while mirror and rotational symmetries were found in both protein in Cyanovirin-N where loops are crystallized in symmetrical conformation the algorithm detects them as a part of symmetrical structure, while in S-Lectin, where they are crystallized in asymmetric conformation they are not detected as parts of larger symmetrical structure, which in this case is mostly compromised of secondary structure elements. This further reinforces the working principle of the automatized symmetry detection showing that it is sensitive enough to distinguish between those fine features simultaneously allowing for detection of large scale symmetries and quasi-symmetries. Lastly we have investigated a structure of more complex monomeric protein triosephosphate isomerase (1IF2) responsible for catalyzing the reversible interconversion between triose phosphates isomers [19], which contains both α-helices and β-sheets intertwined with each other in complex pattern. In this case we detected both mirror symmetry coupled with rotation by 1.7 red around x-axis and trnslation of 3.7nm horizontally, -3.5nm vertically and 5nm in z-direction.

In order to demonstrate the effect of varying bandwidth on scale and exactness of detected symmetries we have tested lens protein γE crystallin (1ZIR) by searching for symmetrical structures within it at varying bandwidths.

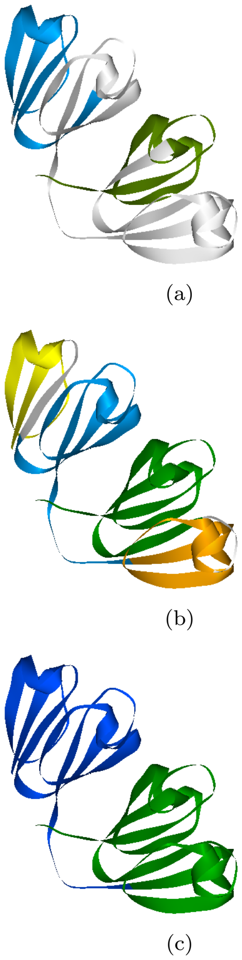

FIG. 2: As bandwidth is varied various subunits are detected as being symmetrical to each other with varying degree of exactness . A. Symmetries in γE crystallin (1ZIR) detected for $h = 0.1$ B. Symmetries detected for $h = 0.2$ C. Global internal symmetry detected for $h = 1.5$

Initially bandwidth $h = 0.1$ corresponding to 10% of the mean distance between all the transformations was used. As shown on Figure 2a this results in detection of symmetry between two small subunits consisting of three β-sheets connected by short α-helix. When the bandwidth is doubled to $h = 0.2$ (Figure 2b) additional two subunits are found and previous subunits are slightly enlarged. Finally when $h = 1.5$ the global symmetry of the protein surfaces.

5. CONCLUSIONS AND FUTURE WORK

In this paper we present theoretical framework that can be applied to obtained automated symmetry prediction algorithm capable of detecting both symmetries and quasisymmetries in protein structures. The algorithm is based upon voting for the existence of different symmetries in the protein and subsequently retrieving meaningful symmetries through clustering and patching. The code was shown to work by detecting symmetries of small protein which can be confirmed by visual inspection. Moreover we were able to show that the degree of exactness required in symmetry detection can be varied leading to detection of larger protein subunits related by quasi- symmetry.

Availability of computerized symmetry detection scheme which can be applied to protein structures can allow for systematic study of symmetries present in proteins and perhaps elucidate new symmetries by performing large scale studies on structures available in PDB. Since evolutionary role of symmetry has been widely hypothesized [1] it would be instructive to search for symmetries and degree of their exactness across phylogenic tree. Moreover new insights into the problem of role of symmetry in protein function can be studied by describing symmetries present in various classes of proteins performing distinct cellular function, such as: soluble enzymes, transcription factors, structural proteins etc. More information about symmetries present in proteins will yield better understanding of role of symmetry in biology.

Acknowledgments

Authors would like to thank Prof. Kardar and Mirny for an interesting topic and Jakub Kominiarczuk for helpful discussion about the mean shift algorithm.

[1] Goodsell DS, Olson AJ, Structural symmetry and protein function, Annu Rev Biophys Biomol Struct. 2000;29:105-53.

[2] Blundell TL, Srinivasan N. 1996. Symmetry, stability, and dynamics of multidomain and multicomponent protein systems. Proc. Natl. Acad. Sci. USA 93:1424348

[3] Cornish-Bowden AJ, Koshland DE Jr. 1971. The quaternary structure of proteins composed of identical subunits. J. Biol. Chem. 246:3092102

[4] Wolynes PG. 1996. Symmetry and the energy landscapes of biomolecules. Proc. Natl. Acad. Sci. USA, 14249-55

[5] Klotz IM. 1967. Protein subunits: a table. Science 155:697-98

[6] Monod J, Wyman J, Changeux J-P. 1965. On the nature of allosteric transitions: a plausible model. J. Mol. Biol. 12:88 118

[7] Ben-Shem A, Frolow F, Nelson N. Evolution of photosystem I - from symmetry through pseudo-symmetry to asymmetry., FEBS Lett. 2004 Apr 30;564(3):274-80.

[8] Vogel C, Morea V., Duplication, divergence and formation of novel protein topologies. Bioessays. 2006 Oct;28(10):973-8.

[9] Grishin, Fold change in evolution of protein structures. J Struct Biol. 2001 May-Jun;134(2-3):167-85.

[10] Andreeva A, Murzin AG., Evolution of protein fold in the presence of functional constraints. Curr Opin Struct Biol. 2006 Jun;16(3):399-408.

[11] Choi IG, Kim SH. Evolution of protein structural classes and protein sequence families. Proc Natl Acad Sci U S A. 2006 Sep 19;103(38):14056-61.

[12] Mitra NL, Guibas LJ, Pauly M, Partial and Approximate Symmetry Detection for 3D Geometry , Siggraph 2006

[13] O. Duda and P. E. Hart, Use of Hough Transformation to Detect Lines and Curves in Pictures, Comm. ACM, 15, 11-15 (1972)

[14] Wang, Xiuyan, Li, Ming, Zheng, Hongyong, Muster, Thomas, Palese, Peter, Beg, Amer A., Garcia-Sastre, Adolfo, Influenza A Virus NS1 Protein Prevents Activation of NF-kappa B and Induction of Alpha/Beta Interferon J. Virol. 2000 74: 11566-11573

[15] Shafer RW, Rhee SY, Pillay D, Miller V, Sandstrom P, Schapiro JM, Kuritzkes DR, Bennett D, HIV-1 protease and reverse transcriptase mutations for drug resistance surveillance. AIDS. 2007 Jan 11;21(2):215-23

[16] M Toba, H Masaki, and T Ohta, Colicin E8, a DNase which indicates an evolutionary relationship between colicins E2 and E3. J Bacteriol. 1988 July; 170(7): 32373242.

[17] Mori T, Boyd MR, Cyanovirin-N, a potent human immunodeficiency virus-inactivating protein, blocks both CD4-dependent and CD4-independent binding of soluble gp120 (sgp120) to target cells, inhibits sCD4-induced binding of sgp120 to cell-associated CXCR4, and dissociates bound sgp120 from target cells., Antimicrob Agents Chemother. 2001 Mar;45(3):664-72.

[18] D Liao, G Kapadia, H Ahmed, GR Vasta, and O Herzberg, Structure of S-Lectin, a Developmentally Regulated Vertebrate β- Galactoside-Binding Protein, PNAS 994;91;1428-1432

[19] Jeremy M. Berg, John L. Tymoczko, and Lubert Stryer, Biochemistry, W. H. Freeman; 6 edition (May 19, 2006)