Supplemental Information: Metric Learning for Enzyme Active-Site Search

Tsuyoshi Kato and Nozomi Nagano

July 17, 2010

Notation

Vectors are denoted by boldfaced lower-case letters, and matrices by boldfaced upper-case letters. Elements of vectors and matrices are not bold-faced. The transpose of a matrix \( A \) is denoted by \( A^\top \); the inverse of \( A \) is denoted by \( A^{-1} \). The \( n \times n \) identity matrix is denoted by \( I_n \). We use \( E_{ij} \) to denote a matrix in which the \((i, j)\) element is one and all others are zero. The \( n \)-dimensional vector, all of whose elements are one, is denoted by \( 1_n \). The \( n \)-dimensional vector, all of whose elements are zero, is denoted by \( 0_n \). Herein, \( \mathbb{R} \) is used to denote the set of real numbers, \( \mathbb{R}^n \) is used to denote the set of \( n \)-dimensional real vectors, and \( \mathbb{R}^{m \times n} \) is used to denote the set of \( m \times n \) real matrices. The set of real nonnegative numbers is denoted by \( \mathbb{R}_+ \). The set of \( n \)-dimensional nonnegative vectors is denoted by \( \mathbb{R}_+^n \). We use \( O_n \) to denote the set of \( n \times n \) rotation matrices. The definition of rotation matrices is given later. The symbol \( \Delta^n \) is used to denote the probability simplex in \( \mathbb{R}^n \) and is defined as

\[ \Delta^n \equiv \left\{ x \in \mathbb{R}_+^n \mid \sum_{i=1}^n x_i = 1 \right\} . \]

Therein, \( \mathbb{N} \) is the set of natural numbers; \( \mathbb{N}_n \) is a subset of \( \mathbb{N} \) defined as \( \mathbb{N}_n \equiv \{ i \in \mathbb{N} \mid i \leq n \} \). Symbols \( \leq \) and \( \geq \) are used to denote not only the standard inequalities between scalars, but also the component-wise inequalities between vectors.

Rigid-Body Transformation

Rigid-body transformation is a class of transformations that consists of a rotation and a translation. The map from point \( x \) in three-dimensional space is expressed as

\[ x \mapsto Rx + v , \]

which describes a rotation determined by the rotation matrix \( R \in O^3 \) followed by a translation determined by translation vector \( v \in \mathbb{R}^3 \). The symbol \( O^3 \) denotes the set of rotation matrices. We say that that square matrix \( R \in \mathbb{R}^{3 \times 3} \) is a rotation matrix if matrix \( R \) is orthonormal (i.e., \( R^\top R = I_3 \)) and the product of the three singular values is positive. If the product of singular values were negative, then the transformation would permit an inappropriate reflection for the comparison of protein structures discussed in this work.

Proof of Theorem 1

Define matrix \( D \in \mathbb{R}^{\ell \times n} \) such that the \((i, j)\)th element is given as

\[ D_{ij} = \| x_{i,j} - \hat{R} x'_{i,j} - \hat{v}_i \|^2 . \]

The matrix clarifies that the constraint for deviations in (3), described in the ‘2.2 Metric learning’ section, is linear as

\[ \text{diag}(y) Dw - \theta y - \xi \leq 0_n . \]
where $y = [y_1, \ldots, y_\ell]^\top$. The operator diag makes a diagonal matrix with $y$ on the diagonal. Let $c$ be an $\ell$-dimensional vector such that the $i$-th element is equal to $1/|I_+|$ if $y_i = +1$; otherwise, $1/|I_-|$. The problem in (3) is thereby reduced to

$$\begin{align*}
\min \quad & c^\top \xi \\
\text{wrt} \quad & \theta \in \mathbb{R}_+, \quad \xi \in \mathbb{R}_+^\ell, \quad w \in \mathbb{R}_+^n, \\
\text{subj to} \quad & \text{diag}(y)Dw - \theta y - \xi \leq 0_n, \\
& 1_\ell^\top w = 1, \quad w \leq C1_\ell.
\end{align*}$$

(4)

This is a linear program. Consequently, the theorem is established.
Table 2: Prediction performance on the template 1acb with different numbers of residues.

| # of residues | Common sites #mtch #mis | AUC EMR MLR | Sensitivity EMR MLR |
|---------------|--------------------------|-------------|--------------------|
| 4             | 557 4,584                | 0.996 0.998 | 0.987 0.998        |
| 3             | 502 4,039                | 0.987 0.998 | 0.958 0.994        |

#mtch and #mis respectively denote the numbers of site matches and mismatches.

Figure 4: Weights for four templates: 1psa (a), 1qk2 (b), 1arg (c), and 3daa (d).
Table 3: Datasets generated using the LSS algorithm. This table lists the number of samples that were detected using the LSS algorithm. In column #mc/#in, the number of inner atoms in the mainchains and the number of all inner atoms are shown. In the column #mc/#out, the number of outer atoms in the mainchains and the number of all outer atoms are indicated; “#mtch” denotes the number of site matches, whereas “#mis” denotes the number of mismatches.

| pdb  | #mc/#inner | #mc/#outer | #mtch | #mis | Reaction type              |
|------|-------------|-------------|-------|------|----------------------------|
| 1acb | 4 / 16      | 4 / 4       | 557   | 6300 | trypsin-type                |
| 1bcs | 4 / 16      | 4 / 4       | 7     | 3001 | Serine carboxypeptidase 2-type |
| 1tyf | 4 / 16      | 4 / 4       | 5     | 3000 | ATP-dependent Clp protease proteolytic subunit-type |
| 1bls | 4 / 18      | 4 / 10      | 10    | 143  | chelatase-type              |
| 2ace | 6 / 18      | 6 / 7       | 43    | 528  | Proteome component PPI-type |
| 1fnt | 4 / 12      | 4 / 8       | 6     | 2881 | serralysin-type             |
| 1af0 | 0 / 29      | 16 / 18     | 19    | 39   | Zinc metalloprotease atelosin-D-type |
| 1atl | 0 / 22      | 12 / 13     | 8     | 280  | Neutrophil collagenase-type |
| 1kfs | 0 / 23      | 16 / 18     | 11    | 33   | carboxypeptidase-type       |
| 1rpa | 0 / 31      | 0 / 6       | 14    | 8    | acid phosphatase-type       |
| 1vce | 0 / 24      | 0 / 4       | 29    | 13   | RNase-type                  |
| 2dhe | 4 / 36      | 4 / 6       | 13    | 15   | dehalogenase-type 1         |
| 1g12 | 4 / 31      | 4 / 6       | 14    | 23   | dehalogenase-type 2         |
| 1kd1 | 0 / 15      | 0 / 5       | 22    | 1361 | polygalacturonase-type      |
| 2bvw | 0 / 8       | 0 / 0       | 21    | 72809| lysozyme-type               |
| 1qk2 | 0 / 8       | 0 / 0       | 16    | 77359| lysozyme-type               |
| 1bg9 | 0 / 13      | 0 / 0       | 49    | 12973| α-amylase-type              |
| 1fh  | 0 / 13      | 0 / 0       | 48    | 12987| α-amylase-type              |
| 1lsw | 0 / 18      | 0 / 2       | 27    | 864  | xylanase type-A-type        |
| 1emh | 0 / 10      | 0 / 0       | 13    | 32573| Uracil-DNA glycosylase-type |
| 1e59 | 0 / 30      | 0 / 5       | 8     | 4    | 2,3-bisphosphoglycerate-dependent phosphoglycerate mutase-type |
| 1dgy | 7 / 18      | 5 / 8       | 5     | 707  | Adenosine kinase-type       |
| 1gjo | 8 / 25      | 0 / 2       | 14    | 632  | receptor-tyrosine kinase-type |
| 1lzo | 0 / 26      | 4 / 15      | 11    | 22   | adenylyl kinase-type        |
| 1equ,2| 2 / 19      | 2 / 8       | 12    | 104  | Glutaminyl-tRNA synthetase-type2 |
| 1c2t | 0 / 14      | 0 / 0       | 8     | 12020| Phosphoribosylglycinamide formyltransferase-type |
| 1e51 | 0 / 6       | 0 / 4       | 57    | 34908| Delta-aminoimidazolecarboxamide hydratase-type |
| 1h7p | 0 / 8       | 0 / 2       | 48    | 33996| Delta-aminoimidazolecarboxamide hydratase-type |
| 1eh3 | 3 / 14      | 9 / 9       | 58    | 650  | Delta-aminoimidazolecarboxamide hydratase-type |
| 1arg | 4 / 21      | 0 / 0       | 61    | 7676 | aminotransferase-type 14    |
| 1eq7 | 4 / 21      | 0 / 0       | 61    | 7524 | aminotransferase-type 12    |
| 1ahy | 0 / 21      | 4 / 4       | 46    | 2387 | aminotransferase-type 14    |
| 1arg,2| 0 / 21      | 4 / 4       | 43    | 2504 | aminotransferase-type 12    |
| 1map | 0 / 13      | 4 / 4       | 88    | 26986| aminotransferase-type 14    |
| 1ams | 0 / 13      | 4 / 4       | 38    | 29143| aminotransferase-type 12    |
| 1ahg | 0 / 21      | 4 / 4       | 35    | 2413 | aminotransferase-type 14    |
| 4tim | 0 / 15      | 0 / 0       | 64    | 6319 | TIM-type 12                 |
| 6tim | 0 / 16      | 0 / 0       | 63    | 4630 | TIM-type 14                 |
| Method  | EMR | MLR | MLR-CIE | MLR-CE | EMP | MLP | MLP-CIE |     |
|---------|-----|-----|---------|--------|-----|-----|---------|-----|
| 1.000   | 0.997 (0.004) | 0.999 (0.001) | 0.999 (0.001) | 0.999 (0.001) | 0.999 (0.001) | 0.999 (0.001) | 0.999 (0.001) |     |
| 1.000   | 0.997 (0.004) | 0.999 (0.001) | 0.999 (0.001) | 0.999 (0.001) | 0.999 (0.001) | 0.999 (0.001) | 0.999 (0.001) |     |
| 1.000   | 0.997 (0.004) | 0.999 (0.001) | 1.000 (0.000) | 0.999 (0.001) | 0.997 (0.002) | 0.995 (0.001) | 0.995 (0.001) |     |
| 1.000   | 0.997 (0.004) | 0.999 (0.001) | 1.000 (0.000) | 0.999 (0.001) | 0.997 (0.002) | 0.995 (0.001) | 0.995 (0.001) |     |
| 1.000   | 0.997 (0.004) | 0.999 (0.001) | 1.000 (0.000) | 0.999 (0.001) | 0.997 (0.002) | 0.995 (0.001) | 0.995 (0.001) |     |
| 1.000   | 0.997 (0.004) | 0.999 (0.001) | 1.000 (0.000) | 0.999 (0.001) | 0.997 (0.002) | 0.995 (0.001) | 0.995 (0.001) |     |
| 1.000   | 0.997 (0.004) | 0.999 (0.001) | 1.000 (0.000) | 0.999 (0.001) | 0.997 (0.002) | 0.995 (0.001) | 0.995 (0.001) |     |
| 1.000   | 0.997 (0.004) | 0.999 (0.001) | 1.000 (0.000) | 0.999 (0.001) | 0.997 (0.002) | 0.995 (0.001) | 0.995 (0.001) |     |
| 1.000   | 0.997 (0.004) | 0.999 (0.001) | 1.000 (0.000) | 0.999 (0.001) | 0.997 (0.002) | 0.995 (0.001) | 0.995 (0.001) |     |
| 1.000   | 0.997 (0.004) | 0.999 (0.001) | 1.000 (0.000) | 0.999 (0.001) | 0.997 (0.002) | 0.995 (0.001) | 0.995 (0.001) |     |
| 1.000   | 0.997 (0.004) | 0.999 (0.001) | 1.000 (0.000) | 0.999 (0.001) | 0.997 (0.002) | 0.995 (0.001) | 0.995 (0.001) |     |
| 1.000   | 0.997 (0.004) | 0.999 (0.001) | 1.000 (0.000) | 0.999 (0.001) | 0.997 (0.002) | 0.995 (0.001) | 0.995 (0.001) |     |

**Table 4: AUC of ROC curves.**

Prediction performance of seven conditions on 45 templates using an evaluation criterion: AUC. Bold red figures in the table represent the best AUC. Underlined blue figures in the table show that the outer atoms are not significantly different from the best AUC. Templates that contain the outer atoms are shown in blue.
| Method | EMR | MLR | MLR-CI | MLR-CE | EMP | MLP | MLP-CE |
|--------|-----|-----|--------|--------|-----|-----|--------|
| TNRd   | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| TNRd   | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| FNRd   | 0.987 | 0.987 | 0.987 | 0.987 | 0.987 | 0.987 | 0.987 |
| FNRd   | 0.987 | 0.987 | 0.987 | 0.987 | 0.987 | 0.987 | 0.987 |
| FPRd   | 0.013 | 0.013 | 0.013 | 0.013 | 0.013 | 0.013 | 0.013 |
| FPRd   | 0.013 | 0.013 | 0.013 | 0.013 | 0.013 | 0.013 | 0.013 |
| AUCd   | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| AUCd   | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |

Table 5: Sensitivities at specificity 0.95.

Prediction performance of seven conditions on 45 templates using an evaluation criterion: sensitivity. Bold red figures in the table represent the best sensitivity. Underlined blue figures in the table show that the performance is not significantly different from the best sensitivity.
Figure 5: Results of template 1jfh. Weights of template atoms (a) and (b); distributions of unweighted and weighted RMSD (b), (c), (d) and (e); the distributions of distance for each atom (g); and the list of template atoms (h). The weight of each atom obtained by metric learning is shown in (b), but those values obtained without metric learning are shown in (a). For plots (a) and (b), carbon, oxygen, and nitrogen atoms are shown in gray, red, and blue, respectively. Plot (c) portrays the distributions of unweighted RMSD for site matches and mismatches in the training dataset, whereas the distributions of weighted RMSD are shown in (d). Distributions of unweighted/weighted RMSD in the test dataset are shown in (e) and (f), respectively. Box-plot (g) shows the distributions of distances for each atom. In (b)–(g), red and blue bars respectively show site matches and mismatches. Panel (h) describes the atom name, the residue id, and the chain id, and the computed weight for each atom in the template.
Figure 6: Results of template Imap. Weights of template atoms (a) and (b); distributions of unweighted and weighted RMSD (b), (c), (d) and (e); the distributions of distance for each atom (g); and the list of template atoms (h). The weight of each atom obtained by metric learning is shown in (b), but those values obtained without metric learning are shown in (a). For plots (a) and (b), carbon, oxygen, and nitrogen atoms are shown in gray, red, and blue, respectively. Plot (c) shows the distributions of unweighted RMSD for site matches and mismatches in the training dataset, whereas the distributions of weighted RMSD are shown in (d). Distributions of unweighted/weighted RMSD in the test dataset are shown in (e) and (f), respectively. Boxplot (g) shows the distributions of distances for each atom. In (b)–(g), red and blue bars respectively show site matches and mismatches. Panel (h) presents the atom name, the residue id, and the chain id, and the computed weight for each atom in the template.