アルゴリズム 52－6
（1996．7．24）

タンパク質における立体構造のコード化とその解析

多田 淳一, 谷口 文浩, 松田 秀雄, 橋本 昭洋
大阪大学大学院基礎工学研究科物理系専攻情報工学分野

本研究では, タンパク質の立体構造を表す文字列を生成する手法を紹介し, 実際の使用例を挙げることで立体構造の比較や検索における本手法の有用性を示す.

タンパク質はアミノ酸が一本鎖状につながったものである. そこで, 連続した4つのアミノ酸の中心炭素原子の座標を p_{n-2}, p_{n-1}, p_n, p_{n+1} とすると, $\overrightarrow{p_{n-1}p_n}$ と $\overrightarrow{p_np_{n+1}}$ がなす角 θ と, p_{n-2}, p_{n-1}, p_n と p_{n-1}, p_n, p_{n+1} で定義される二つの平面がなす角 τ とから (θ,τ) を球座標として持つようなベクトルを得る. 中心炭素原子間の距離はほぼ一定であることから, タンパク質の構造はこのベクトルの列として表現できる. このベクトルを最も近い正二十面体の法線ベクトルで量子化する. 正二十面体の各面に文字を対応させることにより, 一つのベクトルにつき一つの文字が得られる. 各ベクトルを示す文字をつなげることにより, タンパク質鎖の立体構造を表す文字列が得られる.

立体構造が文字列で表されることによって, 既存の様々な文字列処理のアルゴリズムを用いることが可能となる. 本研究では, タンパク質の部分構造であるモチーフの検索を行なった.

An Encoding Method for Tertiary Protein Structures and Their Analysis

Junichi Tada, Fumihiro Taniguchi,
Hideo Matsuda, Akihiro Hashimoto
Physical Science Course, Graduate School of Engineering Science,
Osaka University

We present an encoding method of tertiary protein structures for exploring similar protein structure. In our method, first describe a protein structure as a sequence of vectors obtained by connecting the central carbon atoms of consecutive amino acids in the protein, then quantizing each vector to twenty representatives by fitting it into one of the normal vectors of the twenty faces in an icosahedron, so that a protein structure is represented by a string of twenty characters. The effectiveness of our method is demonstrated by analyzing a motif (biologically significant partial structure of proteins) using a context-free grammar.

1 はじめに

タンパク質は生体内で様々な機能を果たしているが, その機能は立体構造により決定されると考えられている. タンパク質は多数のアミノ酸が一本鎖状にペプチド結合したポリペプチドであり, 一つのタンパク質を構成するアミノ酸の個数は数十から数百とばらつきがある. このようなアミノ酸の鎖が折り畳まれ, タンパク質の立体構造を決定している.

また, タンパク質は, 一次構造, 二次構造, 三次構造と階層的に定義される構造を持つ. アミノ酸の並びが一次構造であるのに対して, 局所的な範囲でアミノ酸残基が作る規則的な主鎖の繰り返し構造を二次構造と呼ぶ. 二次構造には, らせん状構造をした α ヘリックスや β ストランドと呼ばれる伸展構造が横並びに水素結合をした β シートなどがある [1]. さらに, 複数の二次構造などから作られるまとまった立体構造を三次構造と呼ぶ.

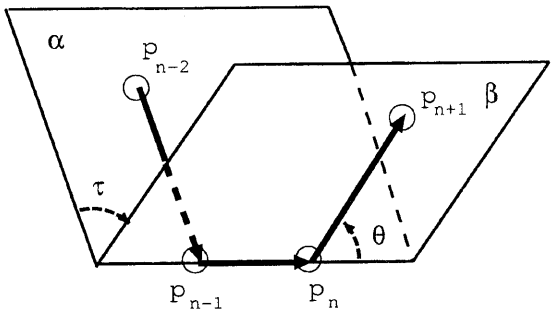


図 1: θ と τ の定義:N-末端側から C-末端側へ連続している 4 つのアミノ酸の C_α 原子の三次元座標をそれぞれ p_{n-2}, p_{n-1}, p_n, p_{n+1} とする.

タンパク質の立体構造は, 分子進化の過程でもよく保存されており, アミノ酸の配列でほとんど相同性がない場合でも, 立体構造は保存されている場合がある. そのため, 立体構造の比較は分子進化の過程を明らかにするのにも有用である.

この様な背景から, タンパク質の立体構造の比較や検索を行ないたいという要求がある. そこで, 我々はタンパク質の立体構造を文字列を使って表現する方法を提案した [2]. タンパク質の立体構造を文字列で表す従来の研究としては, 各アミノ酸での共有結合の周りの回転を表す内部回転角 (二面角ともいう) などの情報から局所立体構造 (α ヘリックスや β ストランドなどの二次構造に対応する) を 7 個程度のグループに分類する方法 [3][4] や, 7 個の連続するアミノ酸の断片を立体構造の類似性により 37 個のグループに分類する方法 [5] がある. これに対して, 我々の方法では後述するようにタンパク質を構成する各アミノ酸の中心炭素原子の座標のみに着目し, それらを結ぶベクトルの列を正二十面体の法線ベクトルで量子化しており, (1) 対象をタンパク質に限定しない汎用的なコード化である, (2) コード化により得られる 20 種類の文字列間の類似度を対応する法線ベクトルの方向余弦で表せるなどの特徴がある.

以下では, この方法の詳細およびその応用として文脈自由文法を使ったモチーフの検索方法とその結果を示す.

2 チェインコード

タンパク質を構成する各アミノ酸は, C_α 原子と呼ばれる構造の中心となる炭素原子を持つ. タンパク質の大まかな構造はこの C_α 原子をつなぐ折れ線で表現することができる.

そこで, この折れ線上の連続する 4 つの原子の座標に着目する. これらの座標を N-末端から C-末端へ順に p_{n-2}, p_{n-1}, p_n, p_{n+1} とする. この時, 3 点 p_{n-2}, p_{n-1}, p_n で決定される平面 α と 3 点 p_{n-1}, p_n, p_{n+1} で決定される平面 β とがなす角をねじれ角 (torsion angle) という [6]. 但し, α を β に重ねる時, 右ねじの進む方向がベクトル $\overrightarrow{p_{n-1}p_n}$ の向きに一致する場合のねじれ角が正であるとする. 本研究では この角度を τ で表すことにする. また, 二つのベクトル $\overrightarrow{p_{n-1}p_n}$ と $\overrightarrow{p_np_{n+1}}$ がなす角を θ で表す. (図 1 参照)

隣合うアミノ酸の距離を一定と考えると, タンパク質の構造は (θ, τ) の配列で表現できる. これを球座標表現された方向ベクトルとみなす. このベクトル (θ, τ) を適当に量子化することにより,

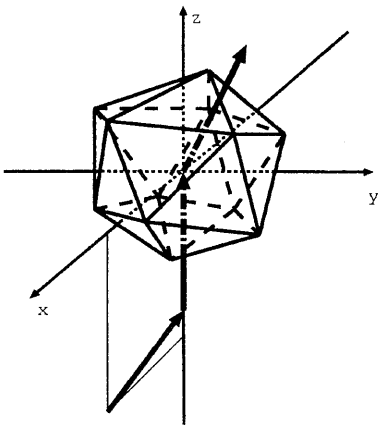


図 2: 重心を原点に一致するように配置された二十面体. 3 つの矢印は p_{n-2}, p_{n-1}, p_n, p_{n+1} を順に結んだもの.

量子化ベクトルの配列が得られる. それぞれの量子化ベクトルに文字を割り振ることにより, 量子化ベクトルの配列から文字列が得られる. このとき, 得られた文字列は平行移動や回転などの座標変換に対して不変である. 本研究では, この文字列を二次元画像処理でのチェインコード [7] との類似性から 3D チェインコード (以下, 単にチェインコードと記す) と呼び, タンパク質の構造の表現方法として用いる. また, 量子化ベクトルとして, 正二十面体の各面の法線ベクトルを用いた.

実際にチェインコードを得るための具体的な手順を以下に示す.
図 1 において,

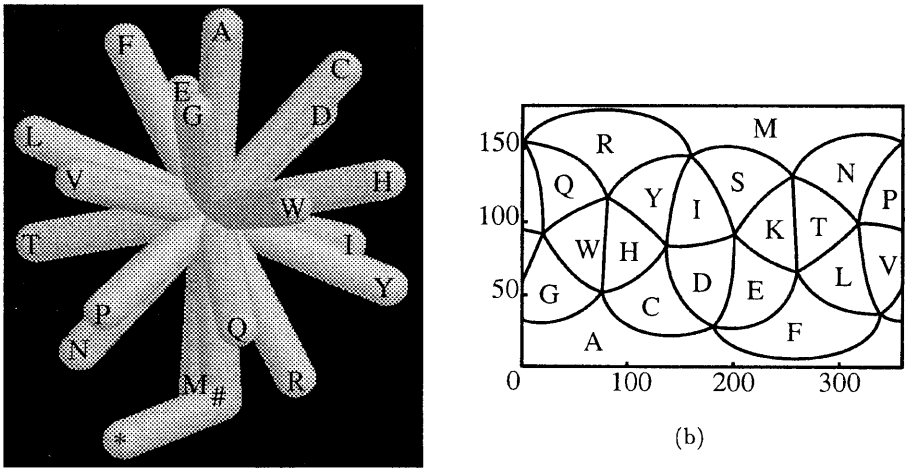
1. p_n を原点に,
2. p_{n-1} を z 座標が負となるような z 軸上に,
3. p_{n-2} を x 座標が正となるような zx 平面上に

それぞれ配置されるように 3 座標を回転平行移動させた時, ベクトル $\overrightarrow{p_np_{n+1}}$ の方向の球座標表現が (θ, τ) となる.

この座標系に, 重心が原点に一致するように正二十面体を配置する (図 2 参照). 配置された正二十面体の法線ベクトルの中でベクトル $\overrightarrow{p_np_{n+1}}$ の方向に最も近いものが, このベクトルに対応する量子化ベクトルとなる. このベクトルを, C_α 原子の座標が p_n であるアミノ酸に対応する量子化ベクトルであるとする.

実際に正二十面体を配置するのには自由度があるが, 本研究では, (θ, τ) 空間におけるヘリックスの分布が正二十面体の一つの面に納まるように考慮し決定した. これにより, チェインコード上でヘリックスの識別が容易になった. 法線ベクトルを 3 次元的に示したものを図 3(a) に示す. また, この正二十面体による (θ, τ) 空間の分割の様子を図 3(b) に示す.

このようにして量子化ベクトルを求める場合, あるアミノ酸に対応する量子化ベクトルを定めるには, N-末端側に二つ, C-末端側に一つのアミノ酸が必要であるため, アミノ酸鎖の N-末端の二つのアミノ酸と C-末端の一つのアミノ酸では量子化ベクトルを定めることができない. しかし,



(a)

(b)

図 3: (a) p_{n-2}, p_{n-1} それぞれが図中の $*$, $\#$ の位置にあり, p_n が放射状の中心にある時, それぞれの法線ベクトルに一致する方向にある p_{n+1} の位置を示した図. 図中示されていないが, S は放射上の中心に対して G と対称な位置に, 同じくK は W と対称な位置にある. (b) 正二十面体による (θ, τ) 空間の分割の様子縦軸が θ の値, 横軸が τ の値, 単位はいずれも (度).

鎖の N-末端の次のアミノ酸については θ を定めることができるので, $\tau = 0$ として, 同様に量子化ベクトルを決定する. これにより, n 個のアミノ酸からなるタンパク質はその構造に対応する $n-2$ 文字のチェインコードで表される.

また, 本研究ではタンパク質の立体構造のデータとして, Protein Data Bank Release 72(以下, PDB-72 と記す) を用いたが, 連続するアミノ酸の C_α 原子間の距離が 7.0 Å を越えるものについては, PDB-72 にそのような記述がなくとも, タンパク質の鎖はそこで切れているものとして扱った.

3 チェインコードの応用

チェインコードにより, タンパク質の立体構造を一つの文字列で表すことができた. よって, 文字列を扱う様々なアルゴリズムをそのまま適用することができる. 特に, チェインコードの部分文字列は, 回転, 平行移動に対して不変であるため, チェインコード同士の文字列照合による比較に分子生物学の分野でよく用いられるアラインメントの手法 [8] を適用できる. チェインコード同士のアラインメントをとることにより, タンパク質の立体構造の類似性を調べることができる. また, 形式言語の理論を用いて, チェインコードの文字列を形式文法の言語としてモデル化し, 構文解析する [9] ことで, タンパク質の立体構造を解析できる.

ここでは, タンパク質の中に見られる局所的構造であるモチーフに注目し, 文脈自由文法の言語としてモデル化することで検索を行なう.

3.1 モチーフとは

類似した機能を持ったタンパク質のグループを解析してみると, しばしば局所的な共通配列パターンが認められる. 一般にこれらのパターンは, 酵素の活性部位など, なんらかの進化的要請から保存されている機能部位である可能性が高い. このような保存配列のパターンをモチーフと呼ぶ. また, モチーフは PROSITE(モチーフデータベース) に正規表現の形で集められている.

以上のことより, 未知配列中に既存のモチーフが存在するかどうかを確かめることは, 未知配列の機能を予測する上でも重要である.

3.2 確率付き文脈自由文法によるモチーフ検索

従来, モチーフの検索にはローカルアラインメントという手法が用いられてきた. しかし, ローカルアラインメントは, 一次構造上の検索であるので, タンパク質の二次構造上の特徴を表現できない. そこで, 文脈自由文法を用いてタンパク質の二次構造上の特徴を生かしながらモチーフを検索することを考える.

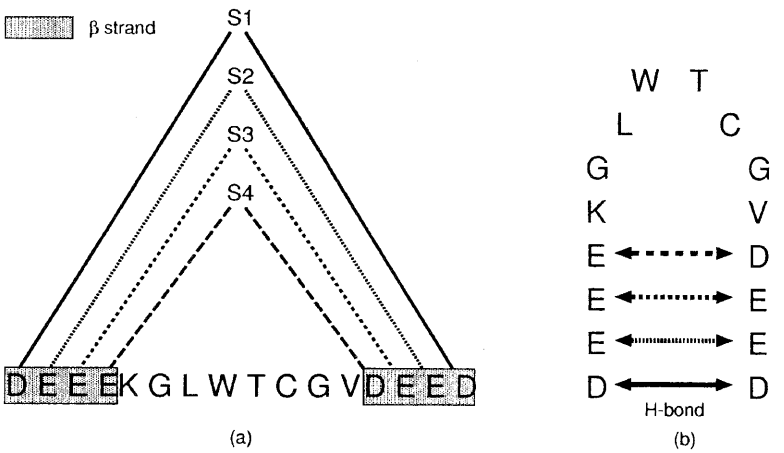


図 4: 構文木 (a) と物理的な構造 (b). (a) で四角で囲まれた部分がストランドであり, S_n は非終端記号を表している. また, S_n からでる二本の枝は水素結合を表している. (b) で矢印はその両端のチェインコードで表されるアミノ酸が水素結合していることを表している.

つまり, モチーフを確率付き文脈自由文法の言語としてモデル化し, チェインコード配列をその文法で構文解析することでモチーフを検索する. 得られた構文木は, 構文上の二次構造であると同時に物理的なタンパク質の二次構造でもある. (図 4 参照) ここでは, 特に β ストランドと呼ばれる伸展構造が横並びに水素結合で連結された β シートと呼ばれる二次構造に注目し, その水素結合を一つの非終端記号からでる二本の枝で表現することを考える. このとき, β シートを作る水素結合は, 次のような生成規則で表される.

$$S_n \Rightarrow A\, S_{n+1}\, B$$

この生成規則は, S_n からチェインコードが A と B であるアミノ酸の水素結合を生成することを表している (S_n, S_{n+1}は非終端記号, A, Bは終端記号である). このように, β シートを作る水素結

合を $S_n \Rightarrow A\,S_{n+1}\,B$ のような生成規則で表すことでタンパク質の二次構造上の特徴を生かしながらモチーフを検索することができる.

ここでは例として, Greek Key(図 5 参照) と呼ばれる二つの β ストランドからなる β シートを持つモチーフを検索することを考える.

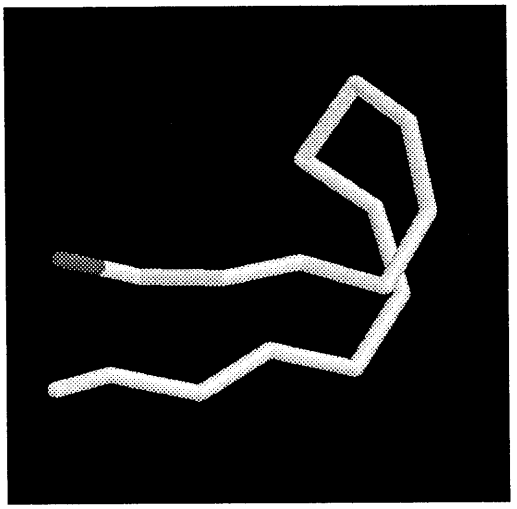


図 5: Greek Key の構造. 二つのストランドが水素結合し, シートを形成している.

3.3 検索結果

モチーフデータベースの PROSITE とアミノ酸配列データベースの SWISS-PROT から Greek Key を持つタンパク質のアミノ酸配列を取りだし, PDB-72 からそのタンパク質の立体構造を取りだし, チェインコードを作成した. この例を図 6 に示す. なお, 図 6 において () で囲まれている配列は, PROSITE の正規表現では Greek Key であると判定できないがチェインコードの類似性から Greek Key とみなしたものである. Greek Key を持つタンパク質は通常 4 個の Greek Key を持つ場合が多いが, PROSITE の正規表現ではまれにそのうちの一つを検出できないことがあり, () で囲まれた配列はその検出に失敗した Greek Key であると考えられる.

Greek Key を確率付き文脈自由文法の言語としてモデル化するために, Greek Key を表す配列中の位置ごとに出現する文字の分布を取り, その分布に応じて対応する生成規則に確率をつけた. このとき, β シートを作る水素結合に対応する部分には, 3.2 で説明した生成規則を用いた.

このようにして得られた確率付き文脈自由文法を用いて, 正例 (Greek Key であるとわかっているチェインコード配列) と負例 (Greek Key ではないが, Greek Key とローカルアライメントをした結果, 比較的スコアの高かったチェインコード配列) に対して構文解析を行ない, そのときの生成確率を調べた. ただし, 正例を構文解析する際には, 正例の集合からその配列を取り除いた配列集合を基に新たに生成規則の確率を計算し直した確率付き文脈自由文法を用いて構文解析した. なお, 配列の数は, 正例が 32 個 (PROSITE の正規表現では検出できなかったがチェインコードの類似性から Greek Key であるとみなした配列 3 個を含む), 負例が 52 個である.

	(a)	(b)
4gcr	EEEEKGLWLCGVEEED	WMLYERPNYQGHQYFL
1prs	EEEEEGEWTCGVEEEC	AILYQNDGFAGDQIEV
2bb2	(DEEEKGLWTCGGEDEE)	(WVGYEQANCKGEQFVF)
2gcr	DEEEKGLGTCGVEEED	WMLYEQPNFTGCQYFL
1ppr	EDEEEWFWTCGVEEEC	AILYQNDGFAGDQIEV
1gcs	EEEEKGFWLCGVEDED	WMLYERPNYQGHQYFL
1blbB	(DEEEKGFWTCGVEEEE)	(WVGYEQANCKGEQFVF)
1blbC	(DEEEEWEWTCGGEDED)	(WVGYEQANCKGEQFVF)

図 6: Greek Key を作るチェインコード配列 (a) とアミノ酸配列 (b) の例. ただし, () で囲まれている配列は PROSITE(モチーフデータベース) で Greek Key であるとの記述はないが Greek Key であると考えられる配列である.

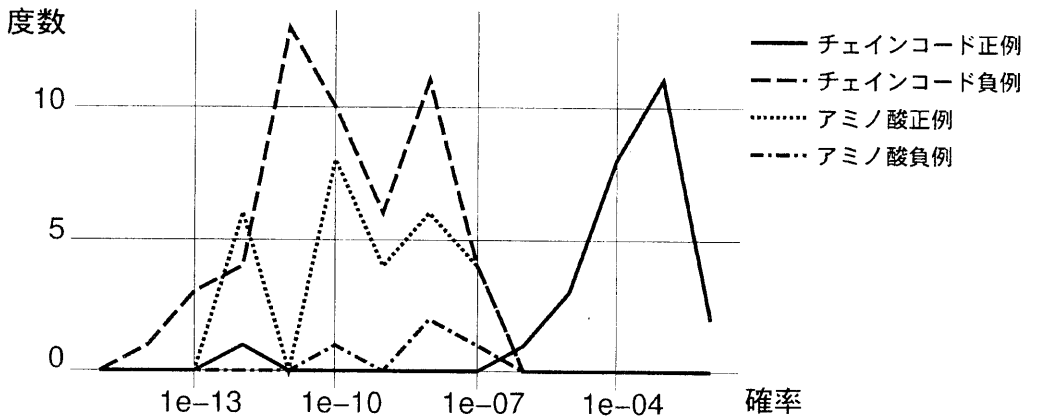


図 7: チェインコード配列, アミノ酸配列の構文解析の結果. 横軸に確率, 縦軸に度数をとっている.

チェインコード配列の正例と負例を構文解析した時の生成確率の度数分布を図 7 に示す. 図 7 では比較のためアミノ酸配列を使って同様の解析をした結果を重ね合わせている.

図 7 からアミノ酸配列を構文解析した場合には, 正例と負例の間に生成確率における差が認められず, 正例と負例を分ける閾値を決定できない. 一方, チェインコード配列を構文解析した場合には, 正例と負例の間にはっきりとした生成確率の差が認められる. 図 7 で, チェインコードの正例と負例の分布曲線の交点を閾値とすると, 正例を Greek Key でないと判断する false negative は, 3.8% であり, 負例を Greek Key であると判断する false positive は, 0% である.

以上の結果からかなり高い認識率で Greek Key を検索できることがわかる.

4 考察

本研究では, タンパク質の立体構造を近似的に表すものとして, チェインコードを提案した. チェインコードを用いることにより, ベクトル列を扱ったり, 座標変換を行なうといった手間をかけずにタンパク質の立体構造の比較ができる. また, チェインコードは文字列によって立体構造を表現しているので, 文字列に対する既存のアルゴリズムをそのまま利用できる.

チェインコードは, ベクトル列を正二十面体の法線ベクトルで量子化するので, 文字列が長いと量子化誤差が累積する恐れがある. しかし, モチーフなどの局所構造を表すチェインコードは比較的短く, 誤差も少ない. つまり, チェインコードはタンパク質の局所構造を比較する上でより有効であると考えられる. 実際, モチーフ検索に対して有効であることは, 本研究でも確認できた.

今回は, タンパク質の二次構造上の特徴である β シートに注目して生成規則を作った. 今後は, より正確な検索を行なうために, 他の特徴である α ヘリックスやターンなども考慮した生成規則を用いたきめの細かい解析を行なう予定である.

参考文献

[1] Chou, P. Y. and Fasman, G. D., "Conformational Parameters for Amino Acids in Helical, Beta-Sheet and Random Coil Regions Calculated from Proteins", *Biochemistry*, vol.13, no.2 pp.211-222, 1974.

[2] Matsuda, H., Taniguchi, F. and Hashimoto, A., "A Notation of Amino Acid Conformations for Exploring Similar Protein Structure", In *Proc. of 1st Pacific Symposium on Biocomputing*, pp.732–733, 1996.

[3] Rooman, M. J., Kocher, J.-P. A. and Wodak, S. J., "Prediction of Protein Backbone Conformation based on 7 Structure Assignments: I Influence of Local Interactions", *J. Mol. Biol.*, vol.221, pp.961–979, 1991.

[4] Miller, R. T., Douthart, R. J. and Dunker, A. K., "An Alphabet of Amino Acid Conformations in Protein", In *Proc. HICSS* , vol.1, pp.689–698, 1993.

[5] Matsuo, Y. and Kanehisa, M., "An Approach to Systematic Detection of Protein Structure Motifs", *CABIOS* , vol.9, no.2, pp.153–159, 1993.

[6] Levitt, M. and Greer, J., "Automatic identification of secondary structure in Globular Proteins", *J. Mol. Biol.*, vol.114 pp.181–239, 1977.

[7] Freeman, H., "On the Encoding of Arbitrary Geometric Configurations", *IRE Trans. Electronic Computer* , vol.EC-10, No.2, pp.260–268, 1961.

[8] Needleman, S.B. and Wunsch, C.D., "A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins", *J. Mol. Biol.*, vol.48, pp.443–453, 1970.

[9] Brown, M. and Wilson, C., "RNA Pseudoknot Modeling Using Intersections of Stochastic Context Free Grammars with Applications to Databese Search", In *Proc. 1st Pacific Symposium on Biocomputing*, pp.109–125, 1996.