博士（理　学）　周　　勇

学 位 論 文 題 名

The development of a software for automating the refinement on the protein structure in X-ray protein crystallography

（X線結晶解析において蛋白質の構造を自動精密化するソフトの開発）

学位論文内容の要旨

X-ray crystallography is the principal method for elucidating the atomic detail of proteins. The basic principle of this method is that the X-ray diffraction data derived from a protein crystal can be considered as the Fourier transform of the electron-density distribution of the crystal. Then through an anti-Fourier transform on the observed diffraction data, an electron-density map can be drawn for the protein crystal. Within the electron-density map the protein structure is reconstructed. From the protein crystal to the final protein structure, several steps are involved, such as diffraction data collection, data processing, electron-density map drawing (phasing), model building, and model refinement. Among these steps, the refinement is the most time-consuming step. The cause of the time consumption is the heavy manual works that have to be done by crystallographers while there have already been many softwares developed for automating or semi-automating the works for other steps. For this reason, a software *Lafire* (Local-correlation-coefficient-based Automatic FItting for REfinement) is developed in this research to automate the refinement of an initial protein model.

In the refinement step, the manual works are performed for the purpose of correcting the errors in the initial model. These errors are derived from the noise in the electron-density map and the resolution limit Usually the manual works are mainly composed of three parts: 1. building the missing parts in the initial model, 2. checking the model and fitting the ill matched parts of the model to the electron density map, 3. refining the model with refinement programs. Performing these works cycle by cycle, a final model can be derived from the initial model. *Lafire* is designed to fulfill these works by programs automatically so that the time taken on the refinement can be reduced.

Corresponding to these sole works done by crystallographers in refining the initial structure model, *Lafire* can be divided into three function modules. The first is for building the missing parts in the model; the second is for fitting the model to the electron density map and the third is for linking the first two modules to the refinement programs and doing the whole process iteratively.

To build the missing parts in the model, an algorithm called MSP (Map Segment Prune) is applied. This algorithm is a simulation of building model by crystallographers from the view of image processing. On the base of the statement that the electron density in the region of protein should appear higher than that in other regions, the map segments for the missing parts are extracted as the continuous high density region between two successive existing fragments. Then the main-chain is traced residue by residue along the map segments. To avoid extending main-chain to the region of side-chain, the side-chain region is separated from the main-chain region through an operation of segment pruning. At last the side-chains are fixed to the built residues with the standard templates designed for each amino acid.

The fitting is performed by *Lafire* through picking out the residues ill matched to the electron density map first, and then adjusting them. To achieve this goal, a grouped local correlation coefficient (GLCC) is proposed as the evaluation of the fitness of residue to the map. By employing the atom density as the element for correlation and applying a modified style for the calculation of correlation, GLCC presents a good efficiency in estimating a residue. Finally the main-chain and the side-chain of the ill matched residues are fitted separately.

At last the generated new model is inputted to the refinement programs. An interface is designed to configure the input files for the refinement programs and commit the whole process from building to refinement iteratively until the evaluation of the model, free R factor, stop declining. Then the result protein model is achieved.

Lafire has been tested on some existing samples, as well as it has been practiced on some new structures. The total amount of these samples is beyond 40. From the results, the conclusion can be drawn that the time consumption on refinement is dramatically reduced by using *Lafire*.

In this dissertation, the detail of the implementation of *Lafire* is introduced. In the last section of the dissertation, the results of the samples refined through *Lafire* are tabulated.

学位論文審査の要旨

主　査　教　授　田　中　　勲
副　査　教　授　龔　　剣　萍
副　査　助教授　渡　邉　信　久
副　査　助教授　芳　賀　　永

学位論文題名

The development of a software for automating the refinement on the protein structure in X-ray protein crystallography

（X線結晶解析において蛋白質の構造を自動精密化するソフトの開発）

構造ゲノム科学プロジェクトの進展により，タンパク質X線結晶構造解析の一連の研究段階が自動化されてきている．これまでにサンプル調製から，回折データ測定，構造解析まで，構造解析の様々な段階を自動化するプログラムシステムが開発され，データ処理，重原子サイトの決定から位相計算・改良，さらにモデリングまでのステップがコンピュータによりほぼ自動的に行われるようになっている．しかし，最新のハイスループット構造解析においても，構造解析の最終段階である結晶構造の精密化は自動化されていない．このため精密化はタンパク質構造解析の全段階で最も時間がかかり，また熟練を要するステップとなっている．本研究では，将来の全自動構造解析プログラムシステム開発に必要不可欠となる，蛋白質構造解析における精密化過程を自動化するためのプログラムの開発を行なった．

タンパク質結晶構造解析の精密化に用いる初期モデルは，得られた初期位相の誤差やデータの分解能の制限などによって，最終構造からかなりずれており，また構造のかなりの部分を欠失しているのが普通である．時として主鎖だけの場合もある．一方でタンパク質の精密化は，回折データと精密化のパラメータの比が低いため，収斂範囲が狭い．これまでに，様々な優れた精密化アルゴリズム（条件付き最小二乗法，エネルギー最小法，分子動力学法，最尤法など）が開発され，またプログラム（SHELX, PLOLSQ, TNT, X-PLOR/CNS, REFMACなど）が作成されてきている．しかし，初期モデルのずれが精密化収斂範囲を超えると，計算だけでモデルを修正することは不可能になる．したがって，精密化の過程では，通常，精密化の計算後にコンピュータグラフィクスを利用して，原子座標を電子密度に合わせる過程，マニュアルフィッティングや，欠失部分を新

しく構築する過程が必要である．本研究では，このようなマニュアル操作をすべて自動的に行なうことで，精密化を全自動化するための，自動精密化プログラム LAFIRE (Local-correlation-coefficient-based Automatic FItting for REfinement) を 開発した．

LAFIRE は未構築部の自動構築，現行モデルの評価，モデルの自動修正機能を持ち，既存精密化プログラムとのインタフェース，精密化ストラテジーとジョブ制御部から成る．プログラム計算には，C, Fortran 言語を，インタフェースとジョブ制御には Shell 言語を使用した．プログラミングと将来の機能の拡張を簡単・明瞭化するという点から，全プログラムの構造はモジュール化した．

Lafire は，既存の自動モデリングプログラム（SOLVE/RESOLVE 或いは ARP/wARP など）で自動的にモデリングできなかった部分，例えばループの構造を，電子密度図とアミノ酸配列に基づいて自動的に構築する機能を持つ．自動構築のためには，画像処理法及び集合理論を利用した新規のアルゴリズム MSP（Map Segment Prune）法を開発した．このアルゴリズムでは，まず未構築部分の電子密度を切り出し，Tree-サーチ法によって，切り出した電子密度から主鎖の流れを確定し，そして，主鎖の構築を行う．その後，標準のテンプレート構造を使って側鎖の構築を行う．LAFIRE のモデルと電子密度図の一致を評価するのには，電子密度図の質を反映する重みをつけた電子密度図の局所的な相関係数 (GLCC) を考案した．電子密度図としては，精密化途中の sigmaa-weight をつけた 2Fo-Fc 電子密度図，あるいは，測定した構造因子 Fo と MAD 位相などから計算した電子密度図を使う．LAFIRE のモデルの修正は，この局所的な相関係数 GLCC により検出した電子密度図と一致しない残基を一つずつ主鎖と側鎖に分けて行う．

開発したプログラムは，これまでに，２０を超える新規なタンパク質の構造精密化に適用され，また Web を介して広く海外にも公開しており，既に１００以上の研究機関によってダウンロードされている．このように Lafire は熟練した研究者のマニュアル作業を必要とした精密化過程を自動的に行なう世界初のプログラムとして機能している．

以上，本研究ではタンパク質の構造解析の中で自動化が最も困難であった精密化の過程に着目し，これを自動化するために新規なアルゴリズムを考案し，実用的なプログラムシステムの作成に成功した． 本研究が生物科学に及ぼす貢献には多大なものがあると考えられ，よって審査員一同は申請者が博士（理学）の学位を得る十分の資格があるものと認めた．