

PC クラスタを用いた大量ゲノムマッピング

理研ゲノム科学総合研究センター
遺伝子構造・機能研究グループ
遺伝子構造機能研究チーム

長谷川 哲

理研ゲノム科学総合研究センター遺伝子構造機能研究チームでは、大量の cDNA (complementary DNA) 塩基配列データが産出され、その解析にゲノムへのマッピングが必要となります。しかしゲノム・マッピングによく用いられている BLAST (Basic Local Alignment Search Tool) は大量の計算機パワーが必要であり、数 CPU 程度の小規模なコンピュータでは限界があります。

本発表ではこのような BLAST を使用した計算の中でも特に、数千万単位の膨大な CAGE (Cap Analysis Gene Expression) タグのゲノム・マッピングにおいて、情報基盤センターの RSCC を使用した例について報告します。

CAGE とは cDNA (mRNA) の 5' 端 20bp 配列を 10 以上連結した 1 つのクローンにすることにより、効率よく生体組織や発生段階ごとの遺伝子の発現パターンや転写開始位置 (TSS: Transcription Start Site) を検出するために当研究室において開発された技術です。

2004 年に行われた FANTOM3 においては合計 2 千万以上の CAGE タグがシーケンスされ、またその後の Genome Network Project においても引き続きシーケンスが行なわれています。このような大量なデータを短期間に計算するために、RSCC を使用し大量のデータを処理する時の工夫や、ゲノム以外の塩基配列データベースのマッピングも円滑に行う為に作った簡単な自動マッピングシステムの説明をします。