「言語寄り道」その①

□意味位相空間で使われている「co-homology」の理論などは、17 世紀のフランスの数学者フェルマーが 360 年前に出した、いわゆる「フェルマーの最終定理」と呼ばれる問題を解決、証明した「中心的理論」です。360 年余りに渡る数学者達の激闘がやっと実ったのが 1995 年で、プリンストン大学で研究をしていたアンドリュー・ワイルズの手によって証明された。この「フェルマーの最終定理」は難攻不落で、数学のあるゆる理論を使い切ってしまうほど難解な問題であった。「楕円曲線論」から始まり、「ガロア表現」そして「代数体の理論」から「円分体論」、「クンマー体論」から「ベルヌーイ数」と「L 関数」を横目に見て、「類数」から「類体論」へ行き、ガロアと合体した「相互法則:フルトブェングラーの定理」を経て、そして日本人が予想した「谷村―志村予想」の証明をして、楕円曲線とモジュラーな楕円曲線を数え、ガロア表現(群)へ変形し、そしてまたまた日本人である「岩澤理論」を使って証明の穴埋めをし、やっと「フェルマーの最終定理」を証明した。ちなみに、「フェルマーの最終定理」とは、高校生でも理解できる、

「$x^n+y^n=z^n$ (ただし $n\geqq 3$)を満たす自然数 x,y,z はない」という問題である。

問題の理解が易しいければ、証明も易しいとは限らない有名な問題です。 意味理解も…そうです。意味解析と意味理解はぜんぜん違いますので宜しく…。

ちょっと「観念」について…。もともと「*idea*」とは、「実念論の祖」と言われたプラトン(427 年～347B.C.)が「実際の個物に対して(物事の理念的本体。思惟されるもの。)がある。」と考えた「普遍主義」「超越主義」とも言われた数学的、神学的「*idea*」論からきている。この人に対して「*eidos*」を唱えたのが、「唯名論の源流」と言われたアリストテレス(384 年～322B.C.)で、「それぞれの個物に内在する *eidos*(エイドス:形相)が本質だ。」と考えた「固体主義」「内在主義」論者である。

意味解析者は、少なくとも「人間を介さなければ意味がない」ので、プラトンを指示するのかな…。

ここで「一言」、「ホモロジー」というとヒューマンゲノムの世界で言われている「ホモロジー検索」と間違える…ヒトもいるかと思い、説明する。

ヒューマンゲノムの「ホモロジー検索」とは…、

1983 年、サルに肉腫をつくるウィルス(SSV)に組み込まれた sis ガン遺伝子産物とヒトの血小板由来増殖因子(PDGF)のアミノ酸配列がそっくりであることが発見された。

この発見には二つの意味で大きな驚きを与えた。

(1)ある程度予測はしていたが、ガン遺伝子が正常な細胞の増殖・分化や固体発生を司る遺伝子そのものであるか、ほとんど同じものであることが判明したこと。

(2)この発見が試験管の中で実験でなく、計算機を用いた「ホモロジー検索」で得られたこと。

これをきっかけに、遺伝子間のもっとも基本的な関係である「相同性(ホモロジー)」を指標とした「ホモロジー検索」が実用的にもっとも重要な遺伝子情報解析法として広く認識されるようになった。

遺伝子の DNA 配列は 4 種類、タンパク質のアミノ酸配列は 20 種類の「文字」で綴られた文章とみなすことができる。

Copyright © 2010 cTag All Rights Reserved.

「文章」を「アライメント」といい「分ち書き」を「ギャップ」という。
「感度」とは「再現率」のことをいい、「選択性」とは「適合率」のことをいう。
ホモロジー検索には、
（1）ダイナミックプログラミング（DP）法
（2）FASTA 法
（3）BLAST 法
などがある。以下順に説明する。
（1）DP 法：
・速度は遅いが、感度と選択性は高い。
・パスカルの 3 角形を用いて、${}_nC_r={}_{n-1}C_{r-1}+{}_{n-1}C_r$という関係を利用すれば、${}_{n'}C_r$（$1 \leqq n' \leqq n$，$0 \leqq r \leqq n'$）の値を定義通りに計算する場合に比べて、はるかに効率よく求めることができる。
・このような手法を変形して、
①表が長方形である
②すぐ左、左上、真上の値も用いる
③足し算の代わりに簡単な演算結果の最大値を求める…などの手法を使う。
・演算は比較する配列の長さの積に比例し、目的の結果に到達できるという最大の特徴がある。
・RNA の2次構造予測やゲノム配列のコード領域予測など遺伝子情報解析分野での応用範囲も広く、比較的並列処理に向いているので専用計算機の開発も進められている。
（2）FASTA 法：
・問い合わせ配列を列行列にし、データベース配列を行行列にして、同じ塩基（アミノ酸）なら格子点に印をつけて、ドットマトリックスを作成し、対角線方向に点が密に並ぶところは局所的に類似性が高い領域を示しているとして、この局所を抽出していく。
・早見表（LookUpTable）を用いると DNA で 1/4 に、アミノ酸で 1/20 の演算量に減らせる。
・k 個の連続した残基を k-タプル（tuple）というが、k-タプル同士が一致する格子点だけを取り出すなら演算回数は $1/4^k$（$1/20^k$）に減らせる。しかし、k をあまり大きく取ると高速になる反面、きめが粗く弱い類似性になる。DNA では k=4～6、タンパク質では k=1～2 が適当とされている。
・早見表を用いた類似領域の高速検索 FASTA 法は、第一段階にすぎず、第 2,3 ステップを経た後、前処理で見つけた類似領域を含む狭い範囲について DP 法を適用する。条件を適切に整えれば DP 法に匹敵する精度と感度を 1/10 の速度で達成する。
・k-タプルの早見表とは、k=1 の場合、与えられた長さ N 個の配列q の各塩基 A,C,G,T を整数配列 T[0…3]の要素の符号化し、位置を保持し、M[1…N]はその位置と同じ種類の塩基が直前に現れた位置を示すことによって、ある塩基の位置すべてがわかるので、類似性の速度が格段に速くなる。
（3）BLAST 法：
・問い合わせ配列を N-k+1 個の単語（k-タプル）に分解する。例えば、“BLASTP”を問い合わせ

Copyright © 2010 cTag All Rights Reserved.

配列、k=3 とすると“BLA”,“LAS”,“AST”,“STP”が単語である。アミノ酸配列の 8000 種類の 3-タプルの中で、それぞれの単語とぴったり並べたときの評価値がある閾値を越えるものを平均 50 個ずつ選び、「類似語」と呼ぶことにする。各類似語にはどの位置の単語の類似語であったかの標識が付けられる。この法の心臓部といえるのが「決定性有限オートマトン（DFA）」である。これはある状態である入力があったときに次に進む状態が一意的に決まるというものであり、例えばデータベース配列のアミノ酸を端から一つずつ読み取り、問い合わせ配列に付随するすべての類似語のどれかを見つけ出したときにヒットしたと答える。一つのアミノ酸を読み取るごとに次に移るべき状態がテーブルとしてあらかじめ計算されているので、ただちに次の読み取り態勢に移れるので、かなり速い。（第１版）

⇒ *cTag*＞意味位相空間ページへ

Copyright © 2010 cTag All Rights Reserved.