生命情報科学専門実習 II

「ゲノムアノテーション」

1　ねらい

専門実習 II の後半ではゲノムアノテーションをおこなう。

アノテーション(*annotation*)とは、あるデータに対して関連する情報（メタデータ）を註釈として付与することを指す(ウィキペディア参照)。ここで、ゲノムアノテーションといった場合、あるデータとはゲノムであり、関連する情報とは遺伝子のコード情報となる。通常は専門家が生物学的な知識を総動員させて、ゲノム配列に遺伝子コードの位置情報と機能情報を付加してゆくことになる。具体的な操作としては、ある生物の全ゲノムの塩基配列が決定された際に、「どこに遺伝子がコードされているか(コード情報)、その遺伝子の機能は何か」を、計算機による情報処理で可能な限り明らかにし、それを基に得られた遺伝子コード情報と機能情報の特定結果を精査し、専門家が生物学的な知識に照らし合わせながら知識情報を付加する(キュレーションと呼ぶ)のが一般的である。現在では、広範囲の多数の生物種のゲノム配列が解読されているが、それらの大部分の遺伝子の機能は、実験ではなく、計算機を用いたこのゲノムアノテーションにより特定されている。

この実習では、原核生物ゲノムの塩基配列の「生データ」から出発してアノテーションを完成させることを目指す。実習に使用するゲノム配列は、カムチャッカ半島の高温の温泉でロシアの科学者が採取した高度好熱菌のもので、ある研究機関との共同研究として諸君の先輩が卒業研究でアノテーションを行ってきた未公開のデータである。DNA 塩基配列データベースには未だ登録されていないので、どのような未知の興味深い性質が潜んでいるのかについて、諸君が最初に発見する可能性もある。この実習を行うねらいは以下の 5 つである。

A)　ゲノムアノテーションはバイオインフォマティクスの総合的な知識を要求するので、これを一から行うことでゲノム情報解析に関する実力が付くはずである。学部レベルでこの水準の実習を行っている大学はないと考えられるので、本学の特徴のある実習と言える。

B)　今回諸君が使用するゲノム配列は、諸君の先輩の卒研生以外は、誰もアノテーションをしていない。この実習の結果は、諸君の先輩が行ったアノテーションと答え合わせをした上で、共同研究機関へ報告を行う。共同研究機関の研究者が精査した上で公式のアノテーションとして全世界に公開される予定である。もし先輩が「機能不明」とした遺伝子に対して、ある特定の機能を妥当な根拠をもとに指摘することができれば、諸君のアノテーションの方が採用される。つまりこの実習は最先端の研究活動の実践（実戦）と言える。その意味でも、本学の特徴のある実習である。

C)　ゲノムアノテーション実習では、これまで行ってきた実習・実験と違い「細かい段取り」を用意しない。自分で道筋を組み立てることを覚えよう。これはどんな仕事をするにせよ必要になる。また、以下に説明するように、この実習はチームワークである。協力しながら一つのことを成し遂げる練習をしよう。

D) ゲノムアノテーションでは、同じ操作(例えば相同性検索や配列の切り出し操作等)を多数回繰り返す必要がある。このような繰り返し操作は、プログラムを作成して一括処理するのが望ましい。ある段階までは手作業で繰り返し作業を行った後に、その経験にもとづいて、各グループで独自にプログラムを作成して手作業の部分を減らすように努力しよう。

E) 使用する配列は未公開データであるので、取り扱いに注意し、外部機関へ持ち出さないこと。通常、自分の扱っている研究試料や情報は個人の所有物ではない。これは何処の会社・研究所でも守るべき研究者(職業人)としてのルールであり、これを守ることを自覚しよう。

2 アノテーションにおけるチーム編成

ゲノムアノテーションのような大規模な解析においては、複数のメンバーでチームを編成して完成するのが一般的である。高等動植物の大型ゲノムでは、複数の国の専門家がチームを編成してアノテーションを行う例が多い。

A) 今回の実習では全員を 4 チーム(各チーム 11ー12 人)に分けて競争する。チーム表は別途配布する。

B) 各チームに約 1.3Mb のゲノム配列を配布する。この実習の目的は 2008 年 1 月 25 日の 3 限目までに与えられたゲノムのアノテーションを完成させることである。

C) チーム内の役割分担(リーダーを作るかつくらないか、仕事を分担するかしないか、分担する際のチーム内での相互チェックはどうするか、プログラム開発は誰が担当するか等)はそれぞれのチームで決める。未公開のゲノムを対象にしたアノテーション実習は今回が 2 度目であるが、昨年度の生命情報科学専門実習 II の経験からすると、チーム内の役割分担のデザインがアノテーションの精度や完成度へ大きく影響していた。構成メンバーの特徴を考慮した上で、チームとして完成度の高いアノテーションを行う戦略を立ててほしい。一日目に方針の概要を決定してほしいが、実習の進行状況を考慮した上で、途中で変更することも重要になる。

D) アノテーションの途中経過は、項目 3 で説明する方法で 11/16 より毎週金曜日に提出すること。各チームの進行状況はコースのホームページに掲載する。

E) **各チームともアノテーションの結果を使って項目 4 に示す設問に答える。また結果は 1 月 25 日に相互に比較し、後日に先輩が卒業研究で行ったアノテーションと比較して達成度を教員が評価する。そのアノテーションの結果に加えて、独自に作成したプログラムを用いて手作業の部分を減らす努力も評価する。**

3　道具と結果のまとめ方

この実習で使う情報解析用の道具はこれまでの実習で習ってきたものである。ただし、計算機環境はこの実習 II の前半でトレーニングした LINUX 環境を使う。また、ホモロジー検索などの負荷の大きい計算は Web を使わず、これも前半で自前の計算機にインストールしたプログラムを使う。特定の公的機関の計算機へ、多人数が同一の IP アドレスから同時アクセスをすると、様々な不都合が起こる可能性があるので、今回の実習では大学内の計算機を使用することにした。個人や少人数で行う解析では、Web を用いて公的機関の計算機を使用するほうが最新のデータベースを利用でき、便利である。

以下にそれぞれの道具（準備と使い方は実習前半で行っている）とゲノムアノテーションの関係を簡単に説明しておく。

a）**Gene finding**　（タンパク質をコードする遺伝子の探索）

まずゲノム配列のどこにタンパク質遺伝子があるのかを探さなくてはならない。これには原核生物やウイルスゲノム配列からの遺伝子発見で一般的に用いられている **Glimmer Ver3**（http://cbcb.umd.edu/software/glimmer/；今回は学内の計算機にインストールしてあるプログラム）を使用する。

b）tRNA 遺伝子の探索

tRNAScan-SE　（tRNA 遺伝子の探索を行うプログラム）を用いる。

c）**Translation**

遺伝子領域が見つかったら、その塩基配列をアミノ酸に翻訳する。これには **EMBOSS (transeq)** を使う。

d）**Homology search**

公的な国際配列データベースを対象にして、見つけた遺伝子候補のアミノ酸配列を用いた配列相同性検索（ホモロジーサーチ）を行い、どの様な名前のタンパク質遺伝子とアミノ酸レベルでの配列相同性が高いのかを調べる。これは nr データベース（non-redundant：重複の無い配列データベース）を検索対象のデータベースとする **BLASTP** サーチで行う。この実習では、それぞれの遺伝子産物であるタンパク質の立体構造が分かっているのかについてもアノテーションを行う。これは同じく **BLAST** を使って、立体構造が分かっているタンパク質のアミノ酸配列を収録した PDB 配列データベース（**pdbaa**）に対して配列相同性検索を行う。

e) Motif search

見つけた遺伝子の産物であるタンパク質がどのようなモチーフ(多数のタンパク質に共有されている類似性の高い部分的なアミノ酸配列で、タンパク質機能と関係する場合が多い)を持つかは、重要なアノテーション項目である。モチーフを探すために ps_scan を使う。

f) Function annotation

見つけた遺伝子が細胞内でどの様な機能を持つのかを調べて分類する。これには特別なプログラムを使わない。実習室に各種辞典・参考図書を用意するので、それらを用いて遺伝子の名前などに基づいて調べる。Web を用いて、Jabion(日本語バイオポータルサイト)の**遺伝子百科**(http://www.bioportal.jp/Gene_search/search/search.cgi)を用いて調べても良い。

次に、各遺伝子に必要なアノテーション項目を具体的に挙げる。なお、(1)-(7)は必須項目、(8)はオプションの項目である。

(1)… **遺伝子の始まりの塩基番号**

(2)… **遺伝子の終わりの塩基番号**

(3)… **遺伝子の向き**

アノテーションする塩基配列の DNA 鎖にコードされている場合は+、相補鎖にコードされている場合は-とする。

(4)… **遺伝子の名前**

(ヒント：例えば blast サーチで相同性の高い配列が、生物系統的に近い他の生物種で見つかった場合、その遺伝子に xx dehydrogenase と名前が付いていて、他にその遺伝子に似たものがその種ならびにアノテーションを行っているゲノム上に存在しない場合には、その遺伝子がオーソログの関係にあると判断して、そのまま xx dehydrogenase として良い。また、このゲノムがより xx dehydrogenase に似た遺伝子を他に持っていたら、前者の遺伝子は xx dehydrogenase-like protein などとすべきと考えられる。BLAST で似たものが見つからなければ、hypothetical gene とか unknown gene とかいう名前にするのが一般的である。「どの程度のレベルの相同性が見つかった場合に高い相同性と呼び、アノテーションとして採用するのかを判断することが重要である。グループ内で充分に議論し、教員とも相談して、共通的な判断基準を設けておこう。

(5)… **機能(function)分類**

現在、広範なゲノムが決定されており、そのような状況の中、オーソロガスな遺伝子セットを定義することによって、ゲノム間における遺伝子機能の比較や推定を行うことが可能となる。また遺伝子の機能の広がりを理解する上でも遺伝子を機能ごとに分類してゆくことは重要である。現在、微生物ゲノムにてオーソロガス遺伝子セットの構築・提供を行っている DB として、

NCBI での COG (Clusters of Orthorogus Genes,
http://www.ncbi.nlm.nih.gov/COG/)

HAMAP (http://au.expasy.org/sprot/hamap/)

MGBD (http://mbgd.genome.ad.jp/)

などが挙げられる。

今回の実習では、**COG** (Clusters of Orthologous Groups)としてグループ化された機能分類を使用する。**COG** では、細胞内における遺伝子の機能を g_8 ページの機

能分類表に示す 25 種類に分類している。その内のどれに相当するか決める。判断には各種辞典・参考書等を用いる。Web を用いて、Jabion の遺伝子百科(http://www.bioportal.jp/Gene_search/search/search.cgi)を用いて調べても良い。

(6)… **アノテーションの根拠**

blast の結果がトップだった他種の遺伝子の ID と E 値を書く(必須)。機能推定に関して参考にした文献の名前とページ数を書く。アノテーションの根拠とした事柄について説明文も記載しておくとなお良い。

(7)… **類似タンパク質の立体構造があったかどうか**

A: 全体が似ているものがあった(PDB の code を書く)
B: 部分的に似ているものがあった(PDB の code を書く)
C: 似ているものは無かった

(8)… **モチーフの同定**

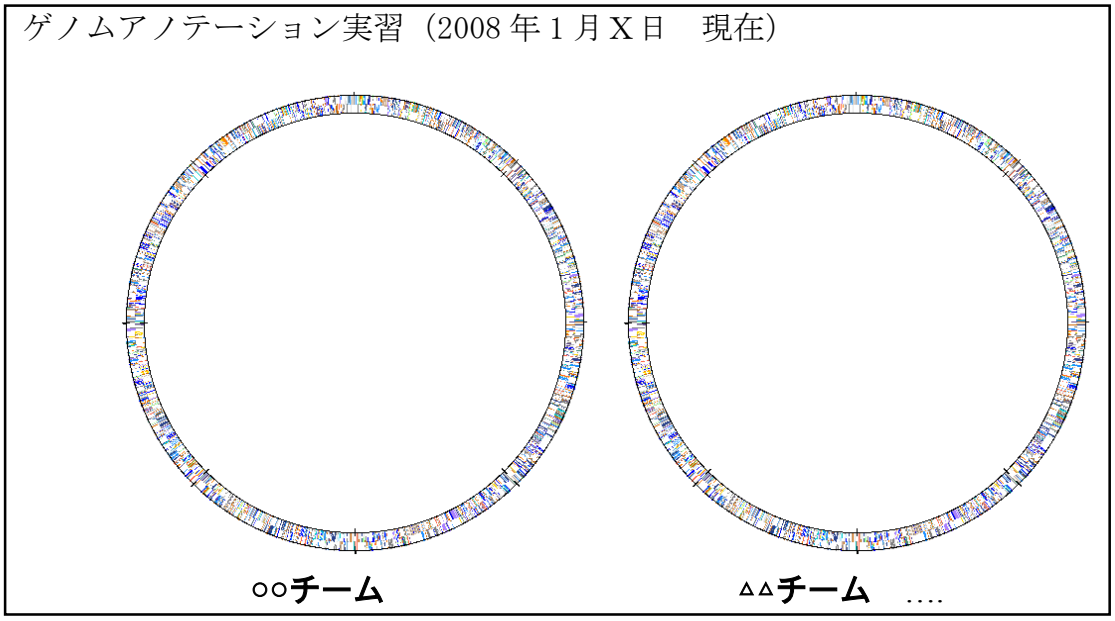
どんなモチーフが見つかったかを記載する。モチーフの ID やモチーフ領域を記載せよ。アノテーションの根拠とした事柄について説明文も記載しておくとなお良い。

アノテーションの結果は以下に例を示す表にし、11 月 16 日より毎週金曜日の実習終了時に提出する。その表をもとにして各グループのアノテーション進捗状況を示すゲノム地図を作成し、コースの HP に掲載していく。提出する表において、遺伝子の始点と終点を glimmer の結果のままにしないこと。また、表の中に２バイト文字（漢字やひらがなど）が含まれないよう注意すること。

図 アノテーション結果の提出例

機能分類表

記号	機能
Information storage and processing (ゲノム複製・転写・翻訳)	
J	Translation, ribosomal structure and biogenesis
A	RNA processing and modification
K	Transcription
L	Replication, recombination and repair
B	Chromatin structure and dynamics
Cellular processes and signaling (細胞の形成・分裂・運動・細胞間のコミュニケーション)	
D	Cell cycle control, cell division, chromosome partitioning
Y	Nuclear structure
V	Defense mechanisms
T	Signal transduction mechanisms
M	Cell wall/membrane/envelope biogenesis
N	Cell motility
Z	Cytoskeleton
W	Extracellular structures
U	Intracellular trafficking, secretion, and vesicular transport
O	Posttranslational modification, protein turnover, chaperones -
Metabolism (代謝)	
C	Energy production and conversion
G	Carbohydrate transport and metabolism
E	Amino acid transport and metabolism
F	Nucleotide transport and metabolism
H	Coenzyme transport and metabolism
I	Lipid transport and metabolism
P	Inorganic ion transport and metabolism
Q	Secondary metabolites biosynthesis, transport and catabolism
Poorly characterized (不明)	
R	General function prediction only
S	Function unknown



毎週提出してもらう表をもとに作成するゲノム地図の例。これを学内 HP に掲示する。なお、このゲノム地図は、GenomeViz というソフトウエアを使って作成する。

4　実習全体の設問　（追加課題等の変更を行うこともある）

以下の問いは、全ゲノムアノテーションを行うことによって始めて答えることができる問いである。この問への回答をグループ単位でまとめ、実習の最終結果として提出してもらう。

（問１）　この微生物はどの生物系統に属するか？　どの微生物に近いと考えられるか？

（問２）　tRNA, rRNA などの非タンパク質遺伝子はどこにどれだけあったか？

（問３）　この生物種は何個の遺伝子(タンパク質だけ)を持つか？

（問４）　この生物種の遺伝子の平均の配列長、遺伝子間距離の平均長、遺伝子コード率(ゲノム配列中で遺伝子がコードされていた割合)は？

（問５）　どの function(機能分類表を参照)の遺伝子がどれだけの比率で存在したか？

（問６）　どれだけの遺伝子(タンパク質)の立体構造が分かっているか？

（問７）　２本鎖 DNA のどちらの鎖に遺伝子が多かったか？

（問８）　コドン使用頻度はどうか？　見つかった各タンパク質遺伝子のコドン使用頻度の全体を集計して、この生物種のコドン使用頻度の特徴を記載せよ。

（問９）　今回のゲノムアノテーションの作業工程の内で、プログラムを作成して処理をするのが適当と思える部分はどの工程か？また、作成したプログラムを説明せよ。

（問１０：オプション問題）　SCOP などのタンパク質立体構造分類を参考にすると、どのタイプの構造のフォールドがどれだけ存在したか？