2020 年度バイオインフォマティクス技術者認定試験
問題解説について

バイオインフォマティクス技術者認定試験は、2020 年度 CBT 方式に移行しました。これに伴い、従来よりも幅広いトピックに関する問題も出題されるようになりました。

この用語解説では、いくつかの新しいトピックや、これまでも出題されていたものの教科書には記載されていないトピックなどについて、簡単な解説を行います。

【生命科学分野】

総論

生命科学分野の問題は、参考書「バイオインフォマティクス入門」（慶応義塾大学出版会）を中心に出題されますが、参考書に掲載されていないもの、例えばエピジェネティクスやゲノム編集のように、比較的新しい重要な項目についても出題されます。ホームページに掲載の出題範囲のキーワードをよく見て準備をしてください。生物は多様で例外も多く、問題文や選択肢を読んでいる中で、厳密には当てはまらないと考えてしまう場合もあるかもしれません。そのような場合は大多数の生物において見られるメカニズムや基本原理を念頭に解答してください。

****

PCR（Polymerase Chain Reaction）

2020 年は COVID-19 の感染拡大により、PCR という言葉が一般に広く浸透した年となった。1983 年に開発され分子生物学に飛躍的な発展をもたらした技術であるが、現在では臨床検査をはじめ様々な分野に応用されている。基本となる反応は DNA 複製の原理と合わせて理解することが必要である。

エピジェネティクス

エピジェネティクスは DNA 塩基配列の変化に依らない、遺伝形質の伝達に関する学問分野である。中心となるのは、DNA のメチル化とヒストンの化学修飾で、これによりヌクレオソームやクロマチンの状態が影響を受け、遺伝子発現が変化することで、同じゲノム配列を持つ細胞でもその形質の違いがもたらされる。また、エピゲノムの状態は後天的な要因により変化し、細胞分裂を通じ継承されうることも知られている。NGS によるエピゲノム解析が発展したことで、バイオインフォマティクスでは重要な応用分野の一つとなっている。

ヒトゲノム・遺伝子解析研究に関する倫理指針

文部科学省、厚生労働省、経済産業省の 3 省が共同で作成している研究倫理指針で、「ヒトゲノム・遺伝子解析研究に関わるすべての関係者においてこの指針を遵守 することが求められる」とあり、バイオインフォマティクス技術者も例外ではない。本指針は研究や社会の変化に合わせて、数年ごとに改定されており、平成 29 年の一部改正では個人情報保護法等の改正に伴って、匿名化や個人識別符号についての項目が改正追加された。内容を暗記する必要はないが、必ず一度は目を通し、基本的な原則を理解することが重要である。

【情報科学分野】

クラウド、SaaS、PaaS、IaaS

近年では Gmail のようにインターネット越しで利用するサービスやツールが増えている。ユーザーが手元にソフトウェアやハードウェアを持たずとも、インターネットを通じてサービスを利用するという形態・考え方はクラウドと呼ばれる。クラウドは、提供されるサービスの種類によって SaaS (Software as a Service、さーす)、PaaS (Platform as a Service、ぱーす)、IaaS (Infrastructure as a Service、いあーす)といった呼び分けがされている。以下にいくつかの例を挙げる。

SaaS	ソフトウェアの提供	Gmail, Slack, LINE, DeepL, Overleaf など
PaaS	開発環境の提供	Google Colaboratory, Microsoft Azure App Service など
IaaS	計算機(サーバー)の提供	Amazon Web Service (AWS) EC2, Microsoft Azure, Google Cloud Compute Engine など

ウェブ版の BLAST（https://blast.ncbi.nlm.nih.gov/）や、UCSC Genome Browser (http://genome.ucsc.edu/）なども SaaS の一種である。なお、実際には SaaS/PaaS/IaaS の境界は曖昧である。たとえばクラウドストレージサービス Dropbox は、単純にユーザーにオンライン上のデータ保存領域を提供する SaaS の一種と考えられるが、ファイルの同期や共有などの制御も可能なため IaaS の一種と考えることもできる。

感度、特異度

新型コロナウイルスの PCR 検査などでも話題になった偽陽性（疑陽性ではない）や偽陰性、感度や特異度といった言葉を整理する。いま、ある疾病のスクリーニング検査を実施した結果が以下の表のように与えられたとすると、検査結果における各指標が計算できる。ただし、以下では *TP, FP, TN, FN* はそれぞれ該当したサンプルの数（人数など）を表すものとする。

	疾病あり	疾病なし	
検査結果が陽性	***TP*** (真陽性, true positive)	***FP*** (偽陽性, false positive)	**陽性的中率=*TP*/(*TP*+*FP*)** (Precision, PPV, 精度, 適合率)
検査結果が陰性	***FN*** (偽陰性, false negative)	***TN*** (真陰性, true negative)	**陰性適中率=*TN*/(*FN*+*TN*)**
	感度＝*TP*/(*TP*+*FN*) (Recall, Sensitivity, 再現率, TPR)	**特異度＝*TN*/(*FP*+*TN*)** (Specificity, Selectivity)	**正確度=(*TP*+*TN*)/(*TP*+*FP*+*TN*+*FN*)** (Accuracy, 精度, 正解率) ※「精度」は Precision との混同に注意
	F-measure=2×Precision×Recall/(Precision+Recall) (F 値, F 尺度, F_1-score)		

深層学習（Deep Learning）

深層学習（ディープラーニング、Deep Learning）とは、ニューラルネットワークを多層に結合した機械学習の一手法である。

【配列分野】

次世代シークエンサデータ解析でよく用いられるデータフォーマット

以下のフォーマットは、次世代シークエンサデータ解析において、様々なコマンドの入出力に使われる標準的なフォーマットなので、解析の際は知っておく必要がある。

- FASTA 形式　配列データを格納する形式として広く用いられており、次世代シークエンサのデータ解析においてはマッピングの対象とするリファレンス配列の形式としてよく用いられる。
- FASTQ 形式　配列データとクオリティ値を合わせて格納する形式で、シークエンサによって読み出された配列（リード）を格納する形式としてよく用いられる。クオリティ値を、アスキーコードに割り当てて 1 文字で表現しており、コンパクトなテキストファイルとなっている。
- SAM 形式　各リードをリファレンス配列上に対応づけ（マッピング）した結果を格納する形式。各リードのリファレンス配列上の位置に加えて、ギャップの挿入などのアラインメント情報もコンパクトな文字列（CIGAR 文字列）によって表現している。
- BAM 形式　テキスト形式である SAM 形式の情報をバイナリ化することで、計算機上で効率よく処理できるようにしたもの。マッピング結果を用いてさらに別のプログラムで解析する場合、BAM 形式に変換して行うことが多い。
- VCF 形式　リファレンス配列上にマッピングされた配列情報に基づいて、変異の検出（バリアントコール）を行った結果を格納する形式。テキスト形式である VCF 形式の情報をバイナリ化した BCF 形式もある。
- GFF 形式　ゲノム配列上の遺伝子（エクソンやイントロン）の位置など、配列上の特徴領域についてのアノテーション情報を格納するための形式。

マルチプルアラインメント

マルチプルアラインメントは、あらかじめスコア体系が与えられていれば、その最適解は（多次元）動的計画法で計算できるが、計算量が配列数に対して指数関数的に増大するために現実的でない。このため、ガイドツリーに沿って類似度の高い配列ペアから順にアラインメントを組み上げる累進法（ツリーベース法）がよく用いられるが、これは一種の貪欲法であり、初期段階で間違ったアラインメントをしてしまうと修正が効かない。これを克服する手法として、作成したアラインメント中の配列をランダムに 2 つのグループに分けて再アラインメントをとる、という処理を繰り返してスコアを改善する手法（逐次改善法）や、配列ペア間でアラインメントを計算する際に最適解以外も候補とし、他の配列とアラインする際の整合性も考慮してアラインメントを決定する手法などがあり、実用的なプログラムではこれらを組み合わせることで精度の改善が

図られている。一方、マルチプルアラインメントにおける配列間の対応は進化過程を反映しているべきであり、アラインメントの評価も進化モデルに基づいて行われるのが望ましい。このため、系統樹に沿った置換、挿入、欠失によってアラインメントが生じるという確率モデルを想定し、これに基づいて系統樹とアラインメントを同時に推定するような方法も開発されている（確率モデルに基づく系統樹推定は、通常は与えられたアラインメント中の置換のみを考慮するが、それと比べて計算が非常に複雑になる）。

BLAST ビットスコア

配列類似性を評価するスコア行列には様々なものがあるが、BLAST ではスコア行列の種類によらず統一的に評価を行えるようにスコアの標準化が行われており、ビットスコアと呼ばれている。ビットスコア S'は、オリジナルのスコア Sから $S'=(\lambda S+\ln K)/\ln 2$ (ln は自然対数)という式により計算される（λと Kはスコア行列によって決まるパラメータ)。このビットスコア S'と E-value の間には、$E=mn2^{-S'}$(m,n は、それぞれクエリ配列、およびデータベース配列全体の長さ)という関係があり、統計的な評価と直接結びついている。ここで 2^{10} がおよそ 1000 であることを考慮すると、ビットスコアが 10 増えるごとに E-value はおよそ 1/1000 になることになり、これを知っておくとスコアを評価する際の目安となる。

【構造分野】

タンパク質デザイン

望みの立体構造や機能を持つタンパク質分子を人工的に設計しようとする試みをタンパク質のデザインと言います。天然由来のタンパク質を部分的に再設計することで生化学的特性をコントロールする研究は古くから行われてきましたが、さらに近年では、コンピュータを用いて機能性タンパク質を文字通りゼロから設計することも可能になってきました。基礎研究の観点からは、タンパク質デザインは立体構造予測の逆問題としても知られており、タンパク質の一次構造と三次構造の間の情報変換原理を理解する上で、両者は互いに対をなす重要なアプローチと言えます。

クライオ電子顕微鏡によるタンパク質立体構造解析

近年、クライオ電子顕微鏡法(クライオ EM)に関するいくつかの技術革新がありました。構造解析法としては、結晶化が不要で比較的短時間で立体構造が得られるなど、様々な利点があります。そのため、PDB では少数派だったこの手法による立体構造データの割合が、急速に増加しています。データを登録する構造生物学者だけでなく、データを利用する側であるインフォマティクスやシミュレーションの立場でも、こういった時流の変化への対応が迫られています。例えば、同じ立体構造データでも部位によって分解能(構造の精度)が大きく異なることがあるなど、クライオ EM データの特性に注意する必要が生じるでしょう。

結晶構造と溶液構造

インフォマティクスでは、極力どのデータも一様に扱いたいものです。PDB に登録されている立体構造データは厳格に定義された書式に従って記述されているので、一見使いやすそうです。しかし実験手法ごとに各種の特性があり、それぞれに注意が必要です。典型的な注意点に、結晶構造と溶液構造の違いがあります。多くの生体分子は、構造的にも、相互作用の観点でも非常にフレキシブルです。結晶構造では、その生体分子が溶液中や細胞中では形成しない構造や相互作用を形成している可能性があります。特に結晶格子中で隣の分子と相互作用している部分には、注意が必要です。溶液構造では、登録されているデータは、非常にバラエティに富んだアンサンブルの単なる代表構造かもしれません。また、電子顕微鏡・X 線結晶構造解析・NMR 解析のいずれの場合も、試料中には存在しているはずの部位が観測できずにモデリングされていない(原子座標が明示されていない)例も多くあります。

DNA の左巻きと右巻き

DNA の二重らせんは、周囲の塩濃度や相対湿度などの条件により、異なる構造をとる

ことが知られている。現在発見されている 6 つの型のうち、生物学的に重要な役割を担うのは図示した A 型 DNA、B 型 DNA、Z 型 DNA とされている。このうち生体内で典型的なのは図の中央に示した B 型である。二重らせん構造をらせん軸上方から見下ろし、5’末端から 3’末端方向にらせん構造をたどると、B 型 DNA では、その軌跡は右巻きとなる。らせん軸に沿って溝幅が広い主溝（メジャーグルーブ)、溝幅が狭い副溝（マイナーグルーブ）が存在する。A 型も B 型同様右巻きである。一方 A 型は、塩基間での結合が緩く、らせん内に隙間が生じた太めのらせん構造であり、主溝や副溝は見られず、塩基対がらせん軸に対して傾いている点が B 型とは異なる。Z 型は、A 型と B 型とは異なり左巻きである。塩基対がらせん軸に対して垂直な点は B 型と同様だが、主溝と副溝の幅にはほとんど差がない。

DNA 二重らせんのイラストを目にする機会は多い。右巻きが描かれている場合、手前側に描かれている鎖が右上から左下に進行する。上述のように、生体内で典型的なのは右巻きである B 型だが、我々が目にするイラストにはどういうわけか左巻きのものが非常に多い。左巻きが描かれる場合、手前側に位置する鎖が左上から右下に向けて進行する。

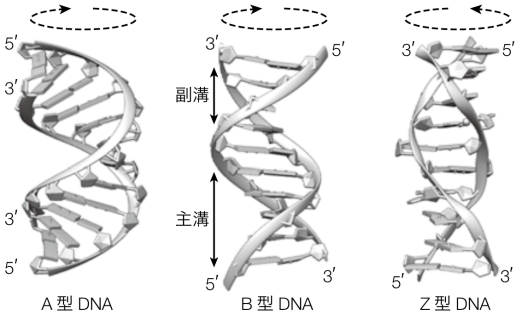


図 異なる DNA 二重らせん構造（バイオインフォマティクス入門 p122 図 1 を抜粋）

ケモインフォマティクス

化学式等で表される化合物の構造や性質を情報科学的に扱う分野をケモインフォマティクスと言います。例えば化学構造式は視覚的に化合物の構造を理解するためには非常に便利ですが、計算機によって機械的に処理するには適していません。大量の化合物に関する情報を計算機によって処理するためには、化合物の構造に関する情報を計算機が処理しやすい形で記述し、またそれを処理する特別な方法が必要となります。ケモインフォマティクスがカバーする領域は多岐に渡りますが、化学構造式を文字列として表現する方法や、化合物の特徴を 0 と 1 のベクトルに変換して類似性を比較する方法などが代表的なものと言えます。

格子モデル

タンパク質の構造を理論的に扱う方法のひとつとして、格子モデルがあります。格子モデルでは二次元ないし三次元の碁盤の目のような格子空間を用意し、アミノ酸残基が格子点一つを占めるものとしてタンパク質を表現します（右図)。1970 年代ごろからよく用いられていたモデルで、タンパク質の構造を大胆に単純化して扱うことで、計算量を削減しつつフォールディング現象を数学的に取り扱うことのできるモデルです。計算機の性能向上も相まって使われる機会は減ってきていますが、タンパク質フォールディングの本質の一端を捉えるためのモデルとして、未だその価値を失っていません。

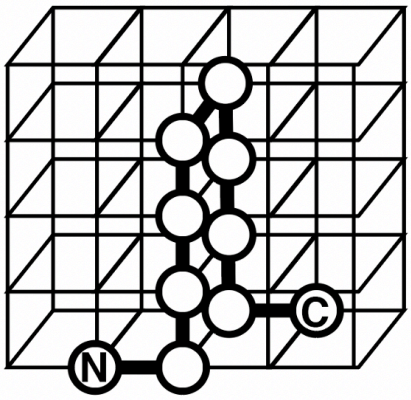


図. 格子モデル。○はアミノ酸残基を示している。格子点上の連続した○でタンパク質を表現する。N、C はそれぞれ N 末端、C 末端を示す。

レビンタール(Levinthal)のパラドクス

これはタンパク質科学では古典的な問題ですが、タンパク質が特定の形に折り畳まれる現象の難しさについて米国の Levinthal が示した考え方です。タンパク質主鎖の ϕ、ψ 角のように、タンパク質を構成する化学結合は回転自由度があるため、全ての角度について適切な値を取ったときのみ正しいフォールドに折り畳まれたと言えます。全ての角度の取り得る角度の組み合わせは天文学的な数に登るため、ひとつひとつのコンフォメーションを高速に試していったとしても正しいフォールドにたどり着くには膨大な時間が掛かるものと見込まれます。しかし実際には、タンパク質はミリ秒オーダーの時間で高速に正しいフォールドを見つけます。

レビンタールのパラドクスは天然のタンパク質が折り畳まれる過程について述べたものですが、この難しさはそのままバイオインフォマティクスにおけるフォールディング予測の難しさにも直結します。高速なアルゴリズムを用いてコンフォメーションの安定性を調べても、その全空間を探索することは難しく、適切に枝刈りして探索空間を狭めることが必要になります。

【遺伝進化分野】

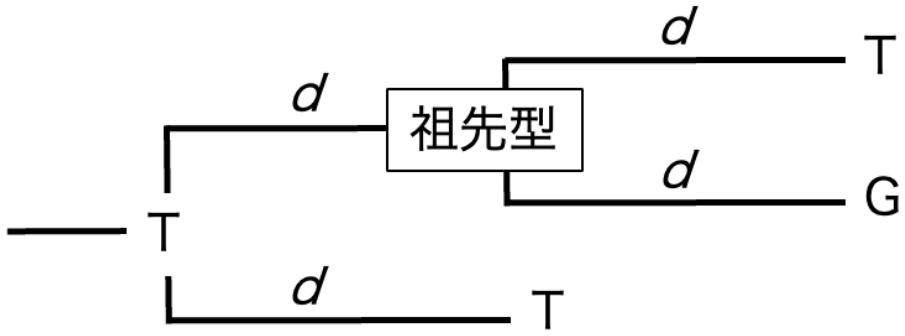
最尤法・祖先遺伝子推定

分子系統樹推定のうち、最尤法は系統樹の樹型・進化距離・現存遺伝子配列と塩基・アミノ酸置換確率に基づいて、系統樹の尤度を評価し、最も尤度の高い系統樹を計算する方法である。この時、尤度計算のために系統樹の内部節(ノード)における塩基・アミノ酸タイプの推定を行うため、この方法は祖先遺伝子配列の推定法としても利用される。

以下の図のような簡単な系統樹(3 つの現存遺伝子からなり、各枝の長さ＝進化距離はすべて dであり、塩基置換確率はすべての塩基の組で等しい場合)を例にとって、尤度計算を考える。葉(端点・リーフ)ノードの T や G が現存遺伝子のある部位の塩基で、祖先型が祖先型塩基(未定)であるとする。化距離 dで塩基置換が起こる確率を pとして、祖先型に A, T, G, C の塩基を当てはめた場合、この系統樹の尤度は以下のようになる。

祖先型 ＝A の場合： p^3（1-p）
祖先型 ＝T の場合： p（1-p）^{3}
祖先型 ＝G の場合： p^2（1-p）^{2}
祖先型 ＝C の場合： p^3（1-p）

進化距離 dの値にもよるが、通常 $p << 1 - p$であることを考えると、必ずしもすべての塩基の組み合わせについて計算する必要はなく、尤度が最も高くなるのは置換回数が少ない場合、すなわち祖先型が T で塩基置換は 1 回だけ起こっている場合であることが分かる。



X 連鎖潜性(劣性)遺伝

ヒト性染色体は XX(女性)または XY(男性)であるが、Y 染色体には遺伝子がほとんどないことから、X 染色体上に潜性(劣性)の疾患関連遺伝子が存在する場合は、女性と男性で表現型の発現の仕方が異なる。女性(XX)の場合は、疾患関連遺伝子をヘテロに持つ場合は保因者となるが発現(罹患)はしないことがほとんどであるのに対して、男性(XY)では一つしかない X 染色体が疾患関連遺伝子を持つ場合は、ほぼ必ず発現(罹患)することになる。逆にいうと、罹患していない男性が保因者であるケースは存在しないと考えて良い。

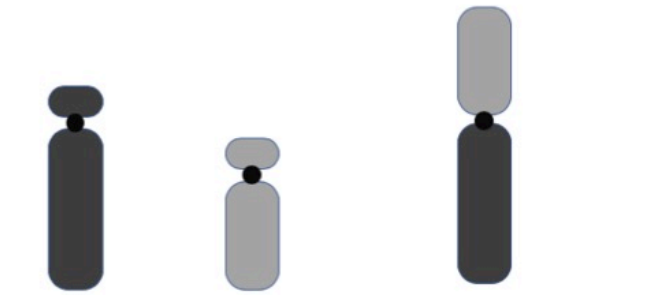
HLA ハプロタイプ

ヒト白血球抗原(human lymphocyte antigen, HLA)、または主要組織適合性複合体(Major Histocompatibility Complex, MHC)と呼ばれる遺伝子は、免疫系における抗原提示に関わり、この遺伝子の対立遺伝子(アリル)の違いが臓器移植・骨髄移植後の拒絶反応の主要因となることが知られている。ヒトゲノムには複数のクラスター化した HLA 遺伝子座(HLA-A, B, C, DRB1 など)が存し、 アリルは HLA—A 遺伝子座では A*24:02、DRB1 遺伝子座では DRB1*15:02 などのように識別され、この組み合わせを HLA ハプロタイプと呼ぶ。臓器移植の際には、適切なドナーを選択するために HLA 型を調査するが、もっとも望ましいドナーはすべてのアリルが完全一致する場合である。

染色体異常

疾患などの原因となる染色体異常には、大きく分けて構造異常と数的異常がある。構造異常は特定の染色体の遺伝子構成が変化する現象であり、欠失(染色体の一部がなくなる)・逆位(染色体の一部が配列的に逆に挿入される)・重複(染色体の一部が複数コピー存在する)・転座(染色体の一部が本来と異なる位置に挿入される)がこれにあたる。数的異常は染色体の本数が変化する現象であり、半数染色体セットの数が異常になる倍数性と、特定の染色体の数が異常になる異数性がある。

21 トリソミー(ダウン症候群)は 21 番染色体が 3 本存在する異数性の例であり、もっとも頻度が高いケースで標準型とも呼ばれる。一方、ロバートソン転座(14;21)は、ヒト 14 番染色体と 21 番染色体の間で、染色体短腕が欠落し、長腕が結合して生じる構造異常(転座)である。ロバートソン転座染色体と正常な 21 番染色体が同一の配偶子に入る場合は、受精により 21 番染色体がトリソミー状態になりダウン症候群の表現型が発現し得る。



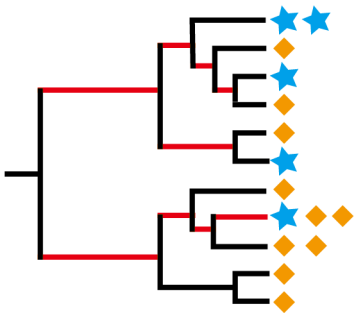
【オーミクス分野】

アンプリコン解析、OTU

メタゲノム解析のように複数ゲノムを一斉に解析する際、特定の遺伝子領域を PCR 増幅して読み出す手法をアンプリコン解析とよぶ。バクテリアの場合は rRNA 遺伝子の 16S サブユニット、真菌の場合はサブユニット間のスペーサー領域（Internal Transcribed Spacer: ITS）を使うことが多い。増幅された配列（アンプリコン）に基づくグルーピング結果は生物種レベルの解像度を持たないため、進化系統樹に同じく OTU（Operational Taxonomy Unit)と呼ばれる。

α多様性とβ多様性、シャノン指標、ユニフラック距離

メタゲノム解析において、同じサンプル内における微生物の多様性をα（アルファ）多様性、異なるサンプル間における多様性をβ（ベータ）多様性と呼ぶ。α多様性の指標としてよく使われるのがシャノン指標で、サンプル内における各生物種の割合のシャノン情報量に相当する。β多様性の指標としてよく使われるのがユニフラック（Unifrac）距離である。二つのサンプルを統合して含まれる種の系統樹を描いた際に、ユニフラック距離とは全枝長に対して片方のサンプルでのみ観測される種に対応する枝長の割合に対応する。つまり 2 つのサンプル間での共通点が少ないほど距離が大きくなる。



まず二つのサンプル（左図、星とひし形）を統合して、含まれる種について系統樹を描くとしよう。その系統樹の枝は、片方のサンプルのみから得られた種のみで構成される枝（赤）と、両方のサンプルで共通に観測された種を含む枝（黒）の二種類に分けられる。このときユニフラック距離とは、前者の枝の長さがすべての枝の長さに占める割合に対応する値となる。

質量分析計と MS/MS

分析化学では質量分析（Mass Spectrometry）を MS と略記し、装置を二つ連結したタンデム質量分析計を MS/MS あるいは MS^2 と書く。この装置では最初の MS で特定の質量をもつ分子のみを選択し、それをフラグメントに分解してから次の MS で測定することで、化合物の部分構造情報を得られる。測定情報をもとに化合物の構造を決定することを同定（identification）と呼ぶ。

精密質量と同位体分布

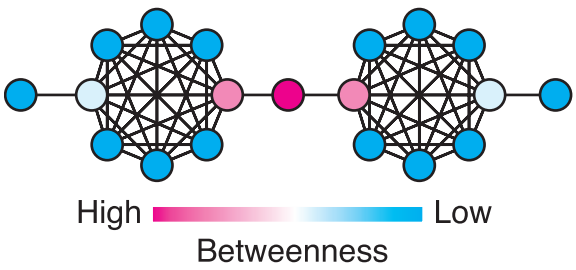
原子番号が同じでも質量数が異なる原子を同位体と呼ぶ。例えば炭素は質量数 12 と 13 のものがそれぞれ 99%と 1%、硫黄は質量数 32 と 34 のものがそれぞれ 95%と 4%、天然に存在する（残り 1%はその他の同位体）。個々の同位体の質量を精密質量と呼び、同位体の分布に基づいた加重平均値を平均質量と呼ぶ。一般の周期表に記載されるのは平均質量だが、メタボロミクスやプロテオミクスで利用する高性能の分析装置では精密質量数を ppm 単位で測定できる。

ハブ、スケールフリー

ネットワークの中で極めて大きいリンク数（あるいは次数）を持つ頂点をハブ（hub）と呼ぶ。頂点に接続するリンク数の統計を取ったとき、リンク数 x をもつ頂点数の割合が $x^{-\lambda}$（λ：定数）というべき分布に比例するネットワークをスケールフリーと呼ぶ。その特徴は、極度にリンクが集中するハブ頂点の存在である。

中心度（centrality）

ネットワークの中で中心となる頂点を選び出すための指標を中心度と呼び、次数（degree）や媒介度（betweenness）が知られる。次数中心度は次数の大きいハブ頂点をネットワークの中心とみなす。媒介中心度は、ネットワーク内の全頂点間の最短経路に最も多く含まれる頂点をネットワークの中心とみなす。この他にも数多くの中心度が提案されており、目的に応じてネットワーク解析に利用される。



左図の例では赤い頂点ほど betweenness が高い。
図は Kasahara et al., PLoS ONE 9(11): e112419 より CC BY4.0 に基づいて転載。

スモールワールド

ネットワークにおいて、任意の 2 頂点間の最短距離の平均値がランダム・ネットワークに近く、それでいて互いに隣接する 3 頂点の割合がランダム・ネットワークよりもはるかに高い場合（いわゆるクラスターをなす場合）、スモールワールドと呼ばれ

る。自然界のネットワークにはスモールワールドが多く見出される（いわゆる友達の友達は友達)。