Computational Characterization of Genomic Sequences

ゲノム配列の数量的特徴付け

By

Chiharu Shioiri

塩入　千春

DISSERTATION

学位論文

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

博士（理学）

Department of Biosystems Science
The Graduate Uneversity for Advanced Studies

総合研究大学院大学　先導科学研究科　生命体科学専攻

平成１４年度

(2002)

親愛なる両親とお世話になった方々へ

要旨

本学位論文は、ゲノム配列の大量データに対する数量的な特徴づけについて述べる。現在、利用可能なミトコンドリアゲノム３５９種、原核生物の染色体１０５配列、シロイヌナズナ、線虫、酵母など６種の真核生物の完全ゲノムのほか、ヒト、マウス、ショウジョウバエのドラフト配列を用いて解析を行った。解析方法は、突然変異の鎖非対称性による塩基組成の偏り（Skew）のパターンを配列全体において調べ、更にDNAの高次構造や機能と関係する二連塩基の頻度傾向を調べた。鎖非対称性はDNAの複製システムと強い相関があり、複製が一方向に進む動物のミトコンドリアでは単調なパターンを持つが、複製が二方向性である細菌では複製の開始点・終止点でSkewが反転し複製開始点の予測も容易に可能である。特に、大腸菌ではグアニン（G）とシトシン（C）との間で非常にきれいなSkewのパターンがあり、それは突然変異のホット・スポットであるChi配列と呼ばれる5'-GCTGGTGG-3'配列がもたらしているようである。真核生物においては、原核生物同様複製の方向は二方向性であるが、一本の染色体上に複製単位が複数あるため、Skewのパターンは複雑である。しかしながら、そのパターンは複製システムと関係していることは明らかであり、リーディング鎖の多い領域とラギング鎖の多い領域など鎖の性質に関係する領域が交互に入れ替わっている可能性が高い。このような一塩基レベルの鎖非対称性にかかわらず、それぞれのゲノム配列全体に一貫した種固有的な二連塩基の頻度傾向がある。原核生物から真核生物まであらゆる種においてTpAは少なく、ほぼ普遍的な傾向である。ヒトや哺乳類の核ゲノムでは、CpG 二連塩基が非常に低く抑えられており、期待値に比べて２０～４０％ほどである。これはCpGのCがメチル化を受けやすいことに起因しており、メチル化を受けたCはデアミ化によってチミン（T）に変化しやすく、C->T転移（トランジッション）が生じて、TpG/CpAになる。そのため、TpG/CpAの増加もみられるが、CpGは極端に抑えられているため、TpG/CpAの増加分だけでCpGの減少分を補填できない。つまり、メチル化-デアミ化による突然変異のほかにCpGを抑制する何らかの機構があると考えられ、最も有力な説はDNAの高次構造と関係するスタッキング・エネルギーが原因とする説である。スタッキング・エネルギーとは原子間あるいは分子間の相互作用による結合エネルギーであるため、それぞれの二連塩基の組み合わせで異なる。そして、このエネルギーによる違いがDNAの高次構造の可塑性を決めるが、TpAの組み合わせは最も可塑性が高くCpGは最も硬質であるうえポリ（CG）は左巻き二重らせんのZ型DNA構造をとる。そのため、TpAやCpGが避けられているとされている。そして、その他の二連塩基も含めて１６通りの組み合わせ全てにおいて、ゲノムに一貫した頻度傾向があることも説

明できる。そして、このような二連塩基の頻度傾向は、アミノ酸をコードしているコード領域でもみられる傾向である。遺伝暗号は４ｘ４ｘ４＝６４通りのコドンからなるが、２０通りのアミノ酸に対応するためコドンには縮退がある。同じアミノ酸をコードする同義コドンの使用頻度は一律ではなくそれぞれの種で偏りがあり、下等な生物ではｔRNA の存在量とも強い相関があるが、ヒトの遺伝子では CpG や TpA を含むコドンの使用頻度が抑えられており、ゲノム全体における二連塩基の頻度傾向の影響を強く受けている。その他の真核生物においても同様に、コドンの使用頻度は二連塩基の頻度傾向の影響を強く受けており、二連塩基の組み合わせは基本的に非常に強い性質であるといえる。

謝辞

本研究を行うにあたり、一貫して先導的かつ懇切丁寧なご指導を頂きました高畑尚之副学長教授に心より感謝を申し上げ、厚く御礼申し上げます。また、審査委員会の先生方、主査の長谷川政美教授、堀内嵩教授、池村淑道教授に深く感謝申し上げます。研究活動におきましてつねに啓発的なご示唆と温かい激励を頂きました先導科学研究科の皆様、特に副指導教官の颯田葉子助教授、そして、森脇和郎前副学長、宝来聰専攻科長、今井弘民教授、笠原正典教授に感謝申し上げます。また、鎖非対称性に関する研究の第一人者でおられるN Sueoka博士には理論的な御教示を頂き、J.R. Lobry 博士には関連論文の御紹介を頂き、感謝申し上げます。

最後に、理化学研究所の職務に就きつつ本論文の執筆を行うことを認めて下さいました、榊佳之プロジェクトリーダーに尊敬と感謝の念を込めて御礼申し上げます。

目次

図の目録

表の目録

１．序章

４５億年前地球が誕生し、３８億年前には最初の生命が誕生した。そのたった一つの原始生命体から現在の多様な生物種が生み出されてきた。このような生物の歴史を生物進化と呼ぶが、歴史や進化を復元することは不可能であるため、その研究には化石あるいはDNA配列といった進化的情報を保有する現存データを科学的に解析する必要がある。進化の研究には主要な課題が二つあり、一つは生物の進化史を明らかにすることであるが、もう一つは進化機構を解明することである。

進化機構の解明は、古典的な集団遺伝学的解析によって為されてきた。長期的進化は本質的には連続した短期的進化の蓄積であるため、分子進化と自然集団の遺伝的多型の維持機構は、同じ進化的現象の異なる側面からの観察である。この遺伝的変異の保有機構について、Kimura(1968)は進化における塩基置換が突然変異と遺伝的浮動によって生じており、集団内の分子レベルでの変異の大部分は、中立かあるいはほぼ中立であると主張した。この中立説は一般に受け入れられ、塩基置換と多型についての興味深い性質も小数のデータから明らかにされてきたものの、一般的なパターンを知るにはより多くのゲノムや遺伝子のデータを解析する必要がある。

Zuckerkandle and Pauling(1965)は、分子レベルでは進化速度が比較的一定であることを発見し、分子時計と呼ばれている。そして、進化の研究が分子レベルで広く行われるようになり、定量的な研究の重要性が認識されてきた。分子進化の基本的な過程はゲノムの大きさとDNA配列の変化にあるので、DNAの進化的変化を定量化するために数学的・統計学的方法が必要となる。そして、近年の国際的なゲノムプロジェクトによる膨大なデータを情報学的に解析する必要性から、生物情報学という新たな一分野も確立されている一方で、Ohno(1970)が提唱したゲノムの進化における遺伝子重複の重要性が、データの蓄積とともにあらためて認識されている。これらのゲノムデータは、遺伝学や進化学だけでなく、古生物学・発生学・分類学などの基礎科学から、医療・農学・遺伝子工学といった応用分野まであらゆる方面に活かされ、各分野の最新の知見は総合的かつ融合的にお互いの分野を啓発し、そして人類共有の知的財産として社会に還元されている。

本研究では、こうしたゲノム配列の基本的な特徴を知るために、現在利用可能なゲノム配列の大量データを用いて、進化学的側面から数量的に特徴付けを行う。第一に、塩基組成の偏りと鎖非対称性を明らかにし、DNAの複製システムとの関係を論じる。第二に、二連塩基の頻度と期待値との比から、ゲノムの固有的特性を明らかにし、更にコドン使用頻度との関係を論じる。最後に、こ

れらの結果から、ゲノムの進化について包括的な考察を行う。

1.1 生物進化と種の多様性

生物の進化史では、種（あるいは属、科、目）を進化の単位とするが、種概念や分類体系に関する議論がこれまでに盛んに行われてきた。種概念に関していまだに明確な定義はないが、種とは自然な状態で交雑可能な個体群と一般に認識されている。種の多様性を記述する方法として、１８世紀 Linne は生物を属名と種名で表す二名法を確立し、分類学を大成した。そして、現在学名のあるものだけでも１５０万種、まだ記載の無いものを含めると地球上に１億種以上の生物が生息していると推定されている。

しかしながら、１９世紀初頭まで、地球上に生息する全ての生物は神による不変の創造物であると広く信じられていた。そのようななか、Lamarck は生物が大規模に絶えず変化していると説き、「動物哲学」を著した。そして、種の多様性から進化機構も論じられるようになったが、重要な進化学的研究は Darwin(1859)の進化論から始まる。彼は、遺伝的変異が生じるメカニズムを知らなかったものの、著書「種の起源」の中で、進化は遺伝的変異の存在のもとに自然淘汰によって生ずると提唱した。１９世紀に至り、Mendel が遺伝の法則を発見し、遺伝の仕組みや遺伝的変異の本質が明らかになると、遺伝的変異が自然突然変異に起因することが示され、Darwin の進化論はネオ・ダーウィニズムあるいは進化の総合説と呼ばれるようになった。その理論的基礎として、Fisher、Wright、Haldane が集団遺伝学の数学理論を発展させた。また、Muller はショウジョウバエの遺伝の研究から、Ｘ線照射による遺伝子の突然変異の誘導を発見し、生じた突然変異が受け継がれることから、自然淘汰とは突然変異型の間で増殖率が異なることであり、遺伝子が生命の基本であると述べた。

遺伝子の化学的本体が DNA であることが明らかになると、分子生物学的手法が１９６０年代なかば頃から進化の研究にも導入されるようになった。そして、生物の系統分類に関する主流は、形態比較から DNA 塩基配列やアミノ酸配列を比較する分子系統学的解析へと移っていった。特に、比較形態学的方法が行えない地球上の生物の初期進化に関する研究において、非常に重要な成果を挙げている。Whittaker(1969)は、あらゆる生物を動物界、植物界、菌界、原生生物界、細菌界に分類する５界説を提唱した。細胞の基本的な構造の違いから、細胞核を持たない細菌類、すなわち原核生物と、動物、植物、菌類、原生生物などの細胞核を持つ生物、すなわち真核生物と２つのグループに大きく分類されるが、Woese and Fox(1977)は第三の生物として古細菌を発見し、従来の細菌類を真正細菌として区別した。古細菌は、単細胞で核を持たず細菌のようであるが、細胞膜の構造や小サブユニット・リボソーム(SrRNA)が真正細菌とも真核生物とも異なる。そして、原始地球上の環境を思わせる嫌気的環境に生息するものが多い。Hori and Osawa(1987)は、５S リボソーム RNA の塩基配列から、古細菌は系統的

に真正細菌より真核生物に近いと主張した。

1.2 ゲノム配列と遺伝子構造

ゲノムとは、生物の個体複製や生命活動に必要な遺伝情報を持つ遺伝子セットのことである。近縁種ではゲノム配列上に遺伝子の並びが保存されており、これをシンテニーと呼ぶが、一般にゲノムサイズや遺伝子数などは極めて多様であり、遺伝情報をRNAに貯えるレトロウイルスのような特殊な例もある。

通常、タンパク質の合成を指定する遺伝情報は、３塩基（トリプレット）を一単位とするコドンが、各々一個のアミノ酸にコードされるDNA塩基配列で構成される。４ｘ４ｘ４＝６４通りのコドンが２０通りのアミノ酸に翻訳されるため、コドンには縮重があり、各々のコドン使用頻度は種によってある一定の偏りがあり、種によっては一般的な遺伝暗号の法則からの変則もみられる。

遺伝子の構造は原核生物と真核生物とで異なり、前者はほとんどがシングルエキソンでゲノム全体もほとんどコード領域であるが、後者はエキソン・イントロン構造を持ち、遺伝子も遺伝子内における遺伝情報を含む領域も不連続である。遺伝子はまずメッセンジャーRNA(mRNA)に転写されるが、そこからスプライシングによりエキソンのみが残る。エキソンの大部分はアミノ酸をコードするコード領域であり、一次転写産物から取り除かれる領域をイントロンと呼ぶ。イントロンの配列は通常GTで始まり、AGで終わる。これをGT-AG則と呼ぶ。イントロンは古細菌や真正細菌の一部でも報告され、エキソン・イントロン構造は原始的なものであり、真正細菌や酵母では急速な増殖に適した生物進化の過程でほとんど失われたと推定されている。また、エキソンの中には、タンパク質中の独立した構造、機能単位をコードしているものがあり、このような構造単位や結合部位、触媒部位などをコードしたエキソンの再編成によって、進化の過程で新しいタンパク質が生み出されてきたというエキソン・シャフリング仮説がある。また、オルターナティブ・スプライシングにより、幾通りかのタンパク質を生成することができる点でも、遺伝子の分断構造は有利であると考えられる。

DNA配列には5'->3'の方向性があり、遺伝子は5'->3'方向に転写されるが、遺伝子の上流5'側には遺伝子発現を制御する領域がある。転写開始を行うポリメラーゼが結合するプロモーター領域や、発現量を制御するエンハンサー領域の他、ヒトなどの高等生物においては多くの遺伝子で更に上流にCpGアイランドと呼ばれる遺伝子の不活性化に関与する領域がある。ヒトゲノムでは、CpG二連塩基の頻度が非常に低く抑えられているが、このCpGアイランドではCpGの頻度が高い。これは、CpGのCがメチル化を受けることと関係しているが、ゲノム配列から検出しやすいため遺伝子予測にも用いられる。

1.3 DNA の構造、複製、修復と突然変異

Chargaff(1950)は、すべての種で塩基組成がアデニン（A）＝チミン（T）、グアニン（G）＝シトシン（C）であることを発見した。この意味は、Watson and Crick(1953)が発見した図のようなDNAの二重らせん構造によって明らかになった（図１.1)。彼らのモデル（B型DNAらせん）は、遺伝情報が保存、転写、複製される仕組みを端的に示している。

その後、DNA結晶のX線解析から、DNAの構造は動的でB型のほかA型やZ型など、さまざまな構造をとりうることがわかった。A型、B型DNAは右巻きらせんで、その繰り返し単位は一塩基である。脱水によってB型からA型への変換が起こり、二重らせんになったRNAヘアピンやRNA-DNA混成体はA型らせん構造をとる。Z型DNAは左巻きらせんで、その繰り返し単位は二連塩基である。Z型DNAは、CGCGやCACAなどのようにプリンとピリミジンが交互に繰り返す領域で形成される。

塩基対の中には２つの塩基が同一平面上にないものもあり、これらはプロペラ・ツイストと呼ばれるずれを持ち、ＤＮＡ鎖に沿った塩基の積み重なりの効果を増進する。このような二重らせんのずれや局所的な構造の変化は、塩基配列によって決まる。DNAの特定の標的塩基配列に結合するタンパク質は、その配列特性を反映する二重らせんの形態から、標的配列の存在を感知すると推測される。また、B型らせんはなめらかに弧を描いて曲がったり、局所的な構造をほとんど変化させずに超らせんを形成したりでき、このような変形の容易さが生物学的に重要なのである（Stryer 1995)。

DNAの複製に関しては、その複製単位や方向性が真核生物や原核生物、ミトコンドリアで各々異なる。大腸菌でのDNA複製は単一の複製開始点（*oriC*）で始まり、両方向へ連続して進む。複製には２０種類以上のものタンパク質が必要であり、DnaAタンパクが*oriC*領域を解きほぐし、ATPをエネルギー源とするヘリカーゼであるDnaBタンパクを導入することで複製フォークをつくる。この複製フォークでは親DNAの二本鎖が両方とも新しいDNA合成の鋳型となる。DNAのラギング鎖（後述）の合成はRNAポリメラーゼの一種のプライマーゼによってつくられた短いRNA鎖をプライマーとして始まる。DNA鎖の一方はリーディング鎖と呼ばれ5'->3'に連続的に合成されるが、もう一方はラギング鎖と呼ばれ１ｋｂ程度の断片、岡崎フラグメントが多数5'->3'に不連続に合成され、全体としては3'->5'に合成される（図1.2)。真核生物においても核ゲノムの複製は二方向性であるが、複製単位は細菌のものに比べて小さく複数ある。高等動物のミトコンドリアゲノムはL鎖とH鎖でそれぞれ一つずつ複製開始点があるが、複製は一方向である。古細菌の複製に関しては不明であるが、複製に関与するポリメラーゼは細菌よりも真核生物のものに似ている。DNAの複製には高

精度な修復機構があるが、合成時の誤りが修復されないと突然変異となる。突然変異は塩基対形成の誤りや塩基の共有結合性修飾、塩基の挿入や欠失、逆位によって生じる。挿入や欠失がコード領域で生じると、塩基配列の読み取り枠を変化させる可能性があり、それらをフレームシフト突然変異と呼ぶ。長い挿入や欠失はおもに不等交叉やDNAの転座によるものであり、転座はトランスポゾンを介して起こるが、一つの塩基対から他の塩基対への置換が最も多い。

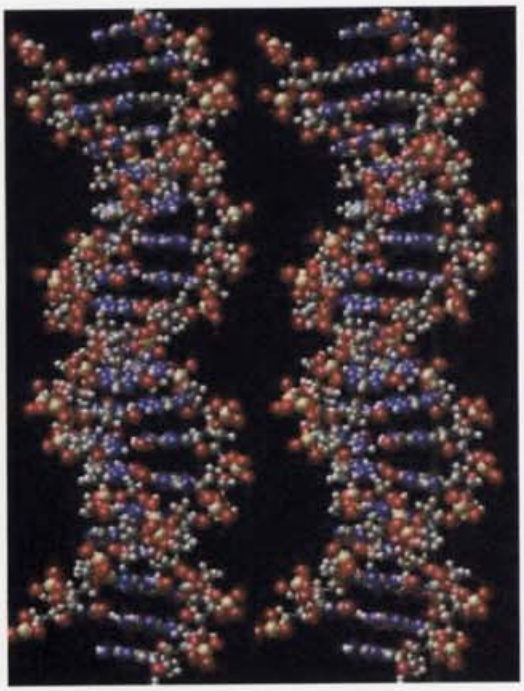


図 1.1　B 型 DNA の二重らせん構造

特徴

1. 逆方向を向いた２本のポリヌクレオチド鎖が共通の軸のまわりにらせんを巻いて、右向きの二重らせんを形成する。
2. プリンとピリミジン塩基はらせんの内側に、リン酸とデオキシリボースは外側にある。
3. アデニン（A）はチミン（T）と、グアニン（G）はシトシン（C）と対合する。
4. 一般的なB型の他、A 型や左向きの Z 型など立体構造にも多型がある。

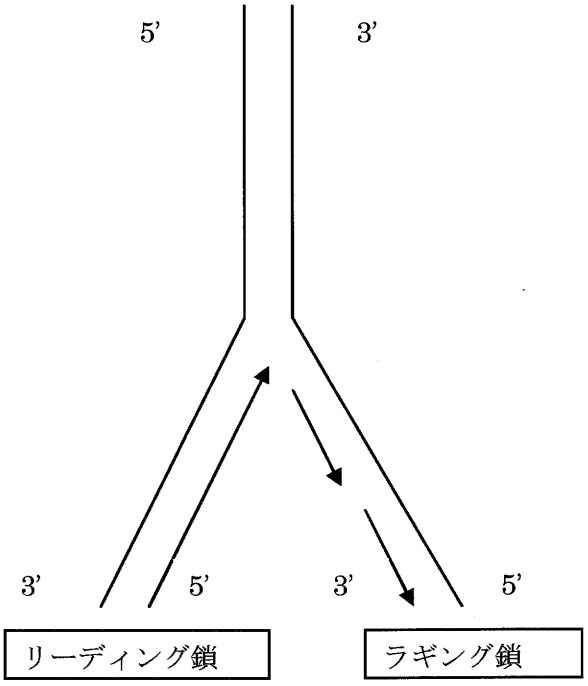


図 1.2 DNA の複製システム。DNA 鎖の一方はリーディング鎖と呼ばれ 5'->3'に連続的に合成されるが、もう一方はラギング鎖と呼ばれ１ｋｂ程度の断片、岡崎フラグメントが多数 5'->3'に不連続に合成され、全体としては 3'->5'に合成される

1.4 塩基置換のパターンと置換速度の推定

置換には２種類あり、プリン塩基同士あるいはピリミジン塩基同士が置き換わるトランジション（転移）と、プリンがピリミジンにあるいはピリミジンがプリンに置き換わるトランスバージョン（転換）と呼ぶ。一般的な傾向として、トランジションの頻度はトランスバージョンに比べて高く、トランジションが自然に発生する機構について Watson and Crick(1953)も触れており、４種の塩基の水素原子のいくつかは位置を変えて互変異性体をつくる可能性があると述べている。アミノ基(-NH_2)が異性化するとイミノ型(=NH)になり、ケト基(-C=O)が異性化するとエノール型(=C-OH)となる。このようなイミノ型やエノール型互変異性体のかたちをとるのは、各塩基についてそれぞれ10^{-4}程度である。互変異性体は一過性であるが、通常の Watson-Crick 対とは異なる塩基対をつくることができる。たとえば、A のイミノ型（A*)互変異性体は C と塩基対を形成し、伸長中の DNA 鎖に A*・C 塩基対が形成されると T が入るはずの位置に C を取り込んでしまう。これが訂正されずに残ると、T->C トランジッションとなる。自然発生突然変異の多くはこうした互変異性体となった塩基による誤った対合から生じる。

置換が４種類の塩基 A, T, C, G の間でランダムに起こる仮定するとトランスバージョンがトランジションより２倍高くなることが予想されるが（図 1.3)、実際には期待される頻度よりもトランジションが多く起こる。しかも、トランジションの４つの型は等頻度ではなく、ゲノムや領域によって異なる。偽遺伝子は機能を失っているため、機能遺伝子に比べて置換速度は速く、例えば哺乳類の偽遺伝子の場合 C->T, G->A 変化の頻度がもっとも高い（図 1.4)。これは、ゲノム配列において CpG 二連塩基の C がメチル化されるとデアミ化によって T に変化しやすいためである。このとき、相補鎖では TpG と相補的に CpA と変化するため、G->A トランジションの頻度も同様に高くなる。前にも述べたが、ゲノム全体の CpG 二連塩基の頻度が非常に低いことはそのためである。しかし、グロビンと ACTH の機能遺伝子の例では、G->A, C->G, A->G が多く、N->G, N->A のようにプリンへの置換が多い（図 1.5)。また、高等動物のミトコンドリアは核の遺伝子に比べ、置換速度が速い。D-loop と呼ばれる制御領域では特に速いため、近縁種や同種内の系統関係を調べるときによく用いられる。図 1.6 のように、ヒトのミトコンドリアの制御領域では、T->C, C->T, G->A, A->G の順にトランジションがトランスバージョンに比べ圧倒的に多い。

塩基置換数を推定する最も基本的な方法は、Jukes-Cantor モデル(1969)である。４種の塩基がそれぞれ等確率で異なる３種の塩基に置換する（すなわち、$\alpha=\beta$）と仮定している。また、トランジション（α）とトランスバージョン（β）を区別した Kimura の２変数法(1980)は、簡略であり広く用いられている。しかし、

現実には上記の例のように4種類の塩基A, T, C, G間の置換速度が置換型により異なるため、様々な置換モデルが提唱され、系統樹解析に用いられている。その多くは、の違いを様々な変数で区別したものであり、現実に進化の過程で塩基組成が変化することまで考慮したものは少ない。Felsenstein モデル(1981)が塩基組成の偏りを考慮しているが、HKY85 モデルはトランジション・トランスバージョンの違いと塩基組成の偏りの両方を考慮している（Hasegawa et al. 1985）。このように、分子データの蓄積とともに、より現実の進化過程に即したモデルが提案されている。

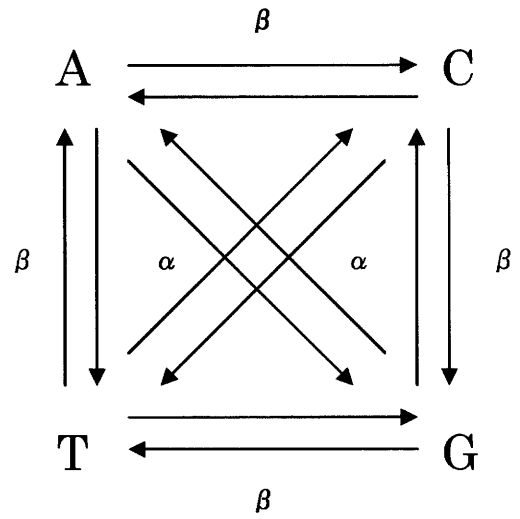
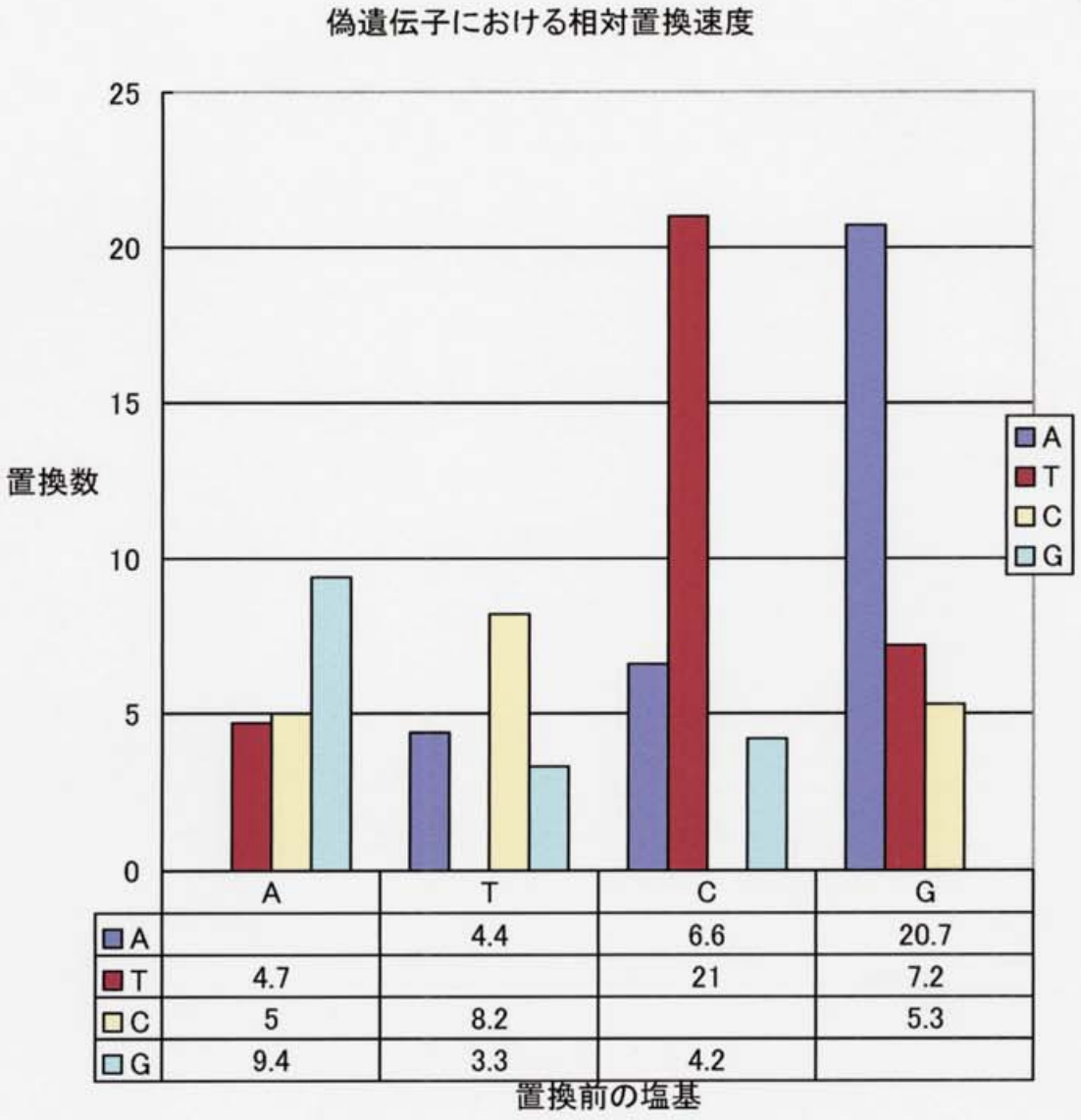
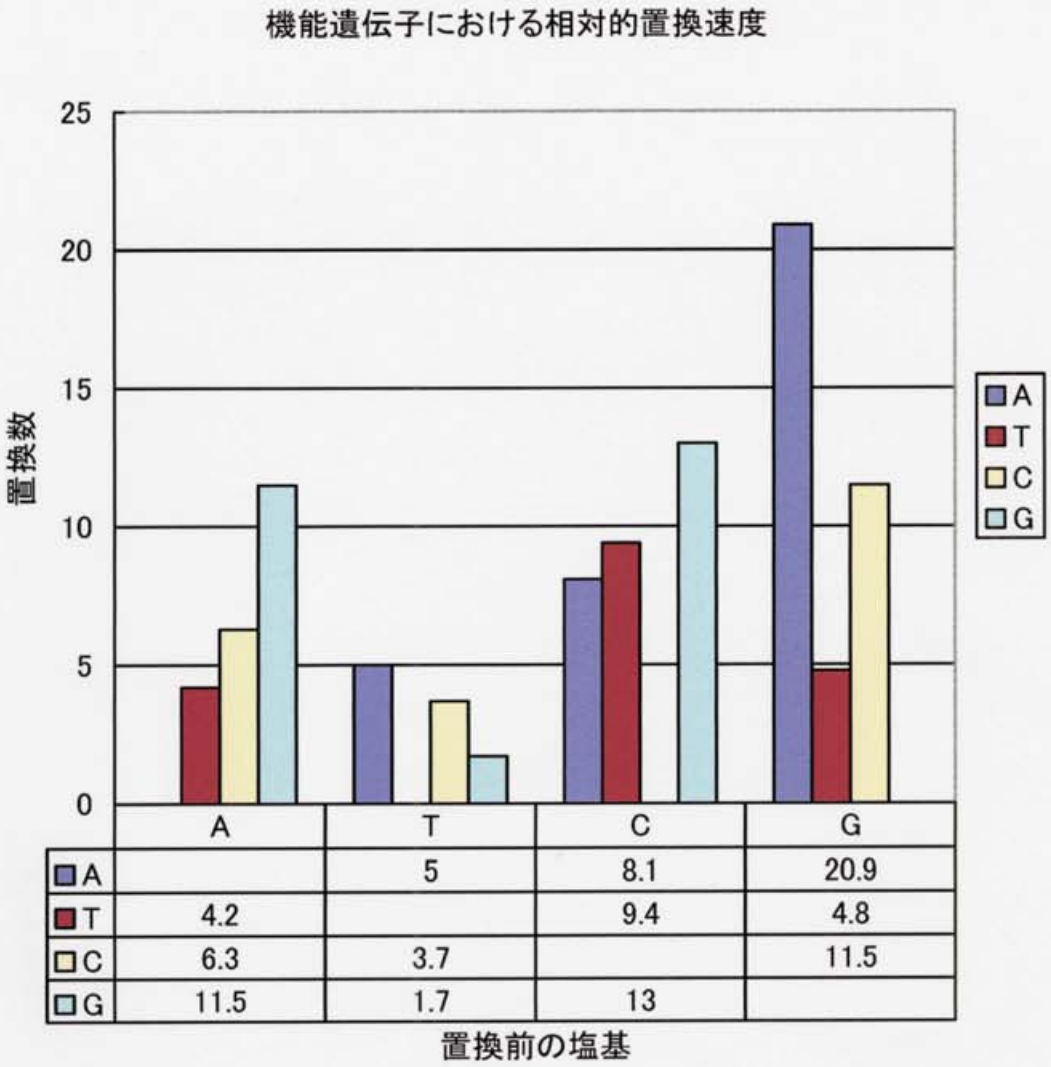


図 1.3 トランジション型塩基置換 **α**（A<->G, T<->C）とトランスバージョン型塩基置換 **β**。



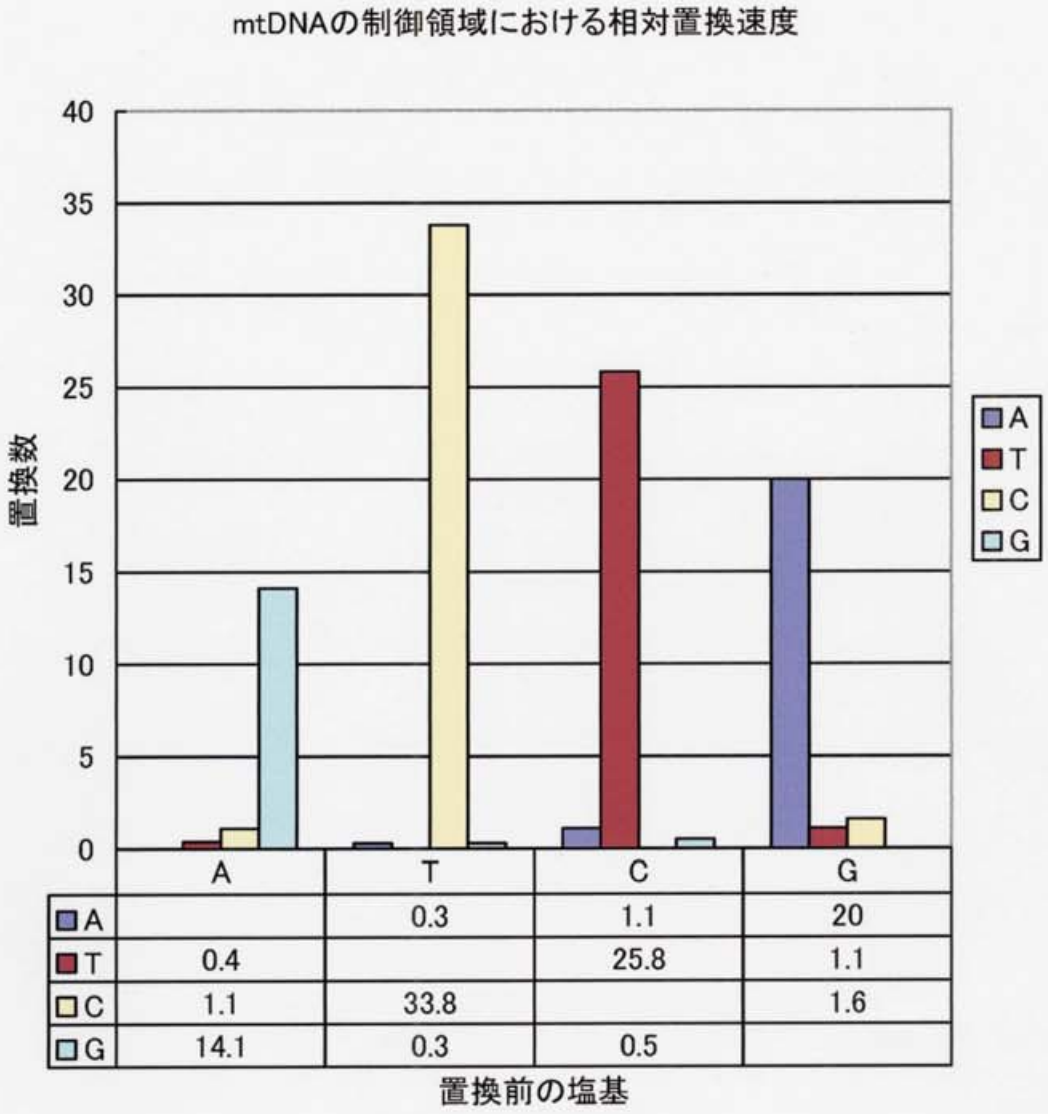
	A	T	C	G
A		4.4	6.6	20.7
T	4.7		21	7.2
C	5	8.2		5.3
G	9.4	3.3	4.2	

図 1.4 偽遺伝子における４種類の塩基 A, T, C, G 間の相対置換速度。これらの値は哺乳類における１６個の偽遺伝子の塩基置換データに基づいている（Li et al. 1984 より）。C->T, G->A トランジションが多い。



	A	T	C	G
A		5	8.1	20.9
T	4.2		9.4	4.8
C	6.3	3.7		11.5
G	11.5	1.7	13	

図 1.5 グロビンと ACTH の機能遺伝子における、4 種類の塩基 A, T, C, G 間の相対的置換速度（Gojobori et al. 1982 より）。G->A, C->G, A->G の順に多いが、N->G, N->A のようにプリンへの置換が多い。



	A	T	C	G
A		0.3	1.1	20
T	0.4		25.8	1.1
C	1.1	33.8		1.6
G	14.1	0.3	0.5	

図 1.6 ヒトの mtDNA の制御領域における 4 種類の塩基 A, T, C, G 間の相対的置換速度（Tamura and Nei 1993 より）。T->C, C->T, G->A, A->G の順にトランジションがトランスバージョンに比べ圧倒的に多い。

1.5 同義および非同義塩基置換

コード領域における塩基置換の多くは、アミノ酸を変化させない同義置換である。同じアミノ酸を指定する複数のコドンは同義コドンと呼ばれ、表 1.1 の遺伝暗号表からわかるようにほとんどの同義コドンは第３ポジションのみが変化している。第１ポジションが変化するものもわずかにあるが、第２ポジションが変化するものはない。したがって、もし突然変異がランダムに生じたとしても、それが集団に固定する塩基置換は、アミノ酸の変化を起こさない第３ポジションに多く検出されるのである（図 1.7)。

コドンにおける塩基サイトを厳密に分類すると、四重縮退、三重縮退、二重縮退、非縮退サイトの４つカテゴリーとなる。四重縮退サイトはすべての塩基変化が同義置換となり、終止コドン(TAA, TAG, TGA)を除く６１通りのセンスコドンうち、３２通りのコドンの第三ポジションである。三重縮退サイトは、３つの可能な変化のうち２つが同義的であり、３通りのイソロイシンコドン(ATT, ATC, ATA)の第三ポジションだけである。二重縮退サイトは、３つの可能な変化のうちの１つだけが同義的であるサイトで、６１通りのセンスコドンのうち２４通りの第三ポジションの他、４通りのロイシンコドン(TTA, TTG, CTA, CTG)と４通りのアルギニンコドン(CGA, CGG, AGA, AGG)の第一ポジションである。残りのすべてのサイトは非縮退である。

コドン間の進化経路を考慮して，同義置換数を推定するための統計学的方法が考案されているが、同義および非同義サイトは DNA 配列において固定されず時間とともに変化するため、塩基置換数の増加とともに推定値の信頼度が減少する。そのため、同義および非同義置換数の相対速度を調べるとき、コドンの３つの塩基ポジションにおける塩基置換数を各々推定することによって、おおよその答えを得ることが多い（図 1.8)。

多くの種や様々な遺伝子で調べられた結果、一般に同義置換速度は非同義置換速度に比べ非常に高いが、両者には有意な相関があり相関係数は 0.55 である（Graur 1985)。置換速度は系統によって大きく異なり、マウスとラットは哺乳類の他の目よりも２倍ほど高く（Wu and Li 1985)、哺乳類のミトコンドリア遺伝子は核遺伝子の１０倍（Brown et al. 1982)、インフルエンザウィルス遺伝子は真核生物の遺伝子より２００万倍も高い（Holland et al. 1982)。核以外の遺伝子で突然変異率が高いのは、DNA あるいは RNA の突然変異による損傷を修復するためのメカニズムが欠如していることによると認識されている（Holland et al. 1982)。

表 1.1 標準遺伝暗号表

コドン	アミノ酸	コドン	アミノ酸	コドン	アミノ酸	コドン	アミノ酸
UUU	**Phe**	**UCU**	**Ser**	**UAU**	**Tyr**	**UGU**	**Cys**
UUC	**Phe**	**UCC**	**Ser**	**UAC**	**Tyr**	**UGC**	**Cys**
UUA	**Leu**	**UCA**	**Ser**	**UAA**	**Stop**	**UGA**	**Stop**
UUG	**Leu**	**UCG**	**Ser**	**UAG**	**Stop**	**UGG**	**Trp**
CUU	**Leu**	**CCU**	**Pro**	**CAU**	**His**	**CGU**	**Arg**
CUC	**Leu**	**CCC**	**Pro**	**CAC**	**His**	**CGC**	**Arg**
CUA	**Leu**	**CCA**	**Pro**	**CAA**	**Gln**	**CGA**	**Arg**
CUG	**Leu**	**CCG**	**Pro**	**CAG**	**Gln**	**CGG**	**Arg**
AUU	**Ile**	**ACU**	**Thr**	**AAU**	**Asn**	**AGU**	**Ser**
AUC	**Ile**	**ACC**	**Thr**	**AAC**	**Asn**	**AGC**	**Ser**
AUA	**Ile**	**ACA**	**Thr**	**AAA**	**Lys**	**AGA**	**Arg**
AUG	**Met**	**ACG**	**Thr**	**AAG**	**Lys**	**AGG**	**Arg**
GUU	**Val**	**GCU**	**Ala**	**GAU**	**Asp**	**GGU**	**Gly**
GUC	**Val**	**GCC**	**Ala**	**GAC**	**Asp**	**GGC**	**Gly**
GUA	**Val**	**GCA**	**Ala**	**GAA**	**Glu**	**GGA**	**Gly**
GUG	**Val**	**GCG**	**Ala**	**GAG**	**Glu**	**GGG**	**Gly**

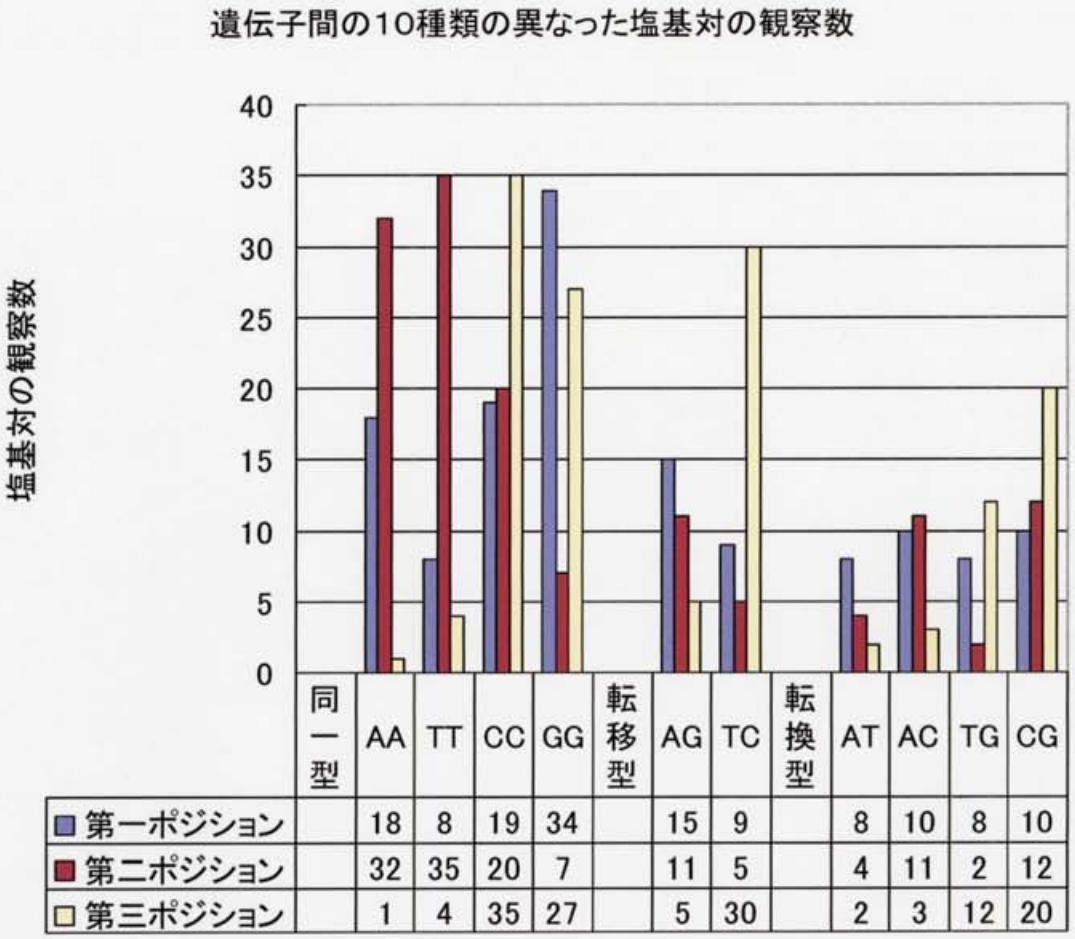
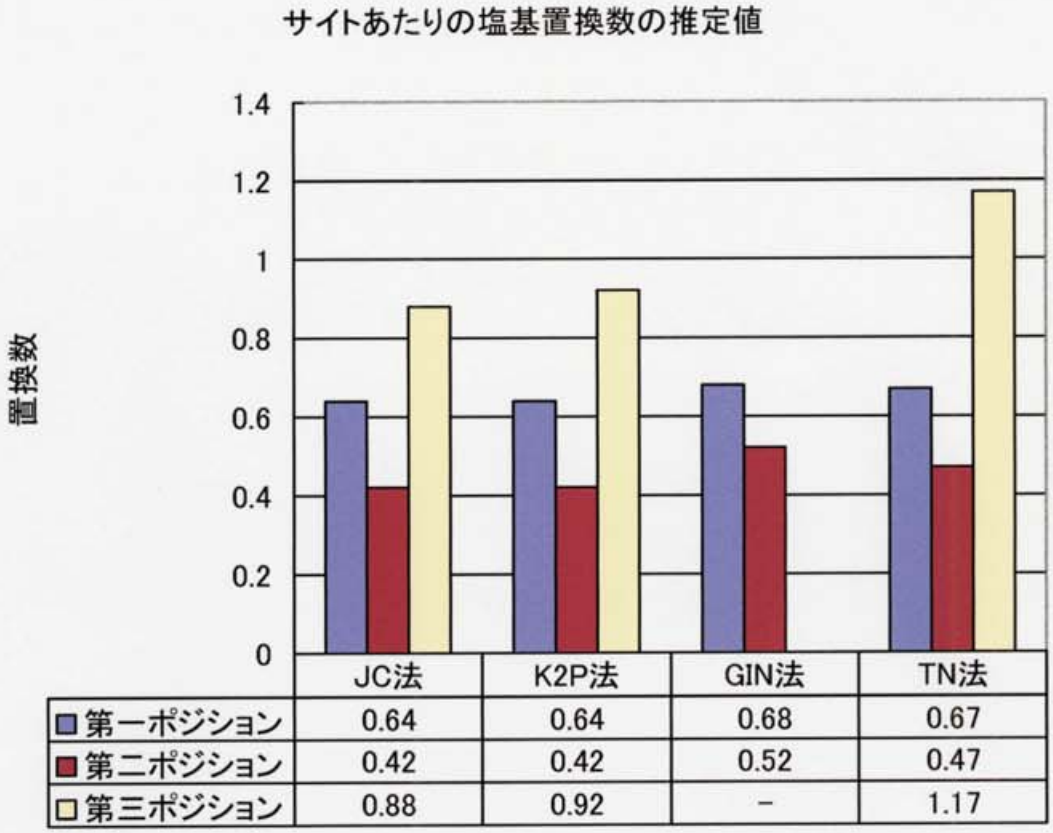


図 1.7　ウサギのαグロビン遺伝子とマウスのβグロビン遺伝子の DNA 配列間における１０種類の異なった塩基対の観察数。コドンのポジションごとの数が別々に示されている（Nei 1987 より）。



	JC法	K2P法	GIN法	TN法
第一ポジション	0.64	0.64	0.68	0.67
第二ポジション	0.42	0.42	0.52	0.47
第三ポジション	0.88	0.92	-	1.17

図 1.8　コドンのポジション別、ウサギのαグロビン遺伝子とマウスのβグロビン遺伝子間の、サイトあたりの塩基置換数の推定値。４種類の推定法による。（JC：Jukes-Cantor.　K2P：Kimura の２変数.　GIN：Gojobori-Ishii-Nei.　TN：Tajima-Nei.　-:適用できない。Nei 1987 より。）

1.6 分子進化と遺伝的多型

分子進化には表現型進化とは違った特徴があるが、いずれも遺伝子の保守的な傾向によるものであり、Kimura and Ohta(1974)は５つの法則にまとめている。

1. 各々のタンパク質について、アミノ酸置換で表した進化速度は、その分子の機能と三次構造が本質的に変わらないかぎり、各種の系統の間でサイトあたり、年あたりほぼ一定である（分子進化時計）。
2. 機能的重要性の低い分子または分子の一部は、重要性の高いものより進化速度が大きい。
3. 既存の分子構造や機能からの変化が小さい保守的な置換は、大きな変化を起こす置換よりも進化の過程で起こりやすい。
4. 新しい機能を持つ遺伝子の出現には、常に遺伝子の重複が先行する。
5. 有害な突然変異遺伝子の自然淘汰による除去と、淘汰に中立または弱有害な突然変異遺伝子の偶然的固定のほうが、有利な突然変異遺伝子に対する正のDarwin 淘汰よりも進化の過程でははるかに頻繁に起こる。

しかし、例外もあり、免疫グロブリンや主要組織適合性抗原複合体（MHC）など、免疫に関する遺伝子は正の淘汰を受けている。MHC は非常に多くの多型があり、５７個のアミノ酸からなる抗原認識領域では、非同義置換速度が同義置換速度の２．５倍と逆転している（Hughes and Nei 1988）。このように、置換速度と遺伝的多型には強い相関があり、実際のデータからも、分子進化と自然集団の遺伝的多型の維持機構が同じ進化現象であることを示唆している。

祖先集団における遺伝的多型、つまり種分化を起こす以前に多型遺伝子の系統はすでに分かれているため（Tajima 1983; Takahata and Nei 1985）、種の系統と遺伝子の系統は厳密には異なる。さらに、核遺伝子とミトコンドリアでは遺伝様式が異なるため、集団の有効な大きさも遺伝子の分岐も異なる。共通祖先の集団の大きさを N とすると、２倍性染色体の核遺伝子では $2N$ 世代であるが、ミトコンドリアでは母性遺伝と半数性とみなせるため、集団の有効な大きさは核遺伝子の 1/4 しかない（Takahata 1985）。従って、ミトコンドリア遺伝子の分岐に比べて核遺伝子の分岐は古くなる。

近年、医療面での重要性から一塩基多型（SNP）のデータが豊富になり、大規模なデータベースも構築されている。このようなデータを進化と多型の維持機構の研究にも活かすことが期待されるが、ヒトの SNP の多くが CpG のサイトであることは、二連塩基の頻度や置換のパターンとも関係しており、極めて興味深い。

２．　材料

ミトコンドリアの完全ゲノム（３５９，表 A.1)、古細菌の完全ゲノム（１５，表 A.2)、細菌の完全ゲノム（９０，表 A.3)、および９種の真核生物(表 A.4)の完全ゲノム（6）とドラフト（3）の DNA 塩基配列データを NCBI の WWW からダウンロードし解析に用いた。

3. 塩基組成の偏りと DNA の複製システム

3.1 背景

Watoson-Crick 対の組み合わせから、一般に A と T、G と C の組成はほぼ等しいとされるが、G・C 対は A・T 対より結合力が強く両者の組成は異なる。G と C の組成を合わせて GC 含量と呼び、一般的な目安にされている。ゲノムにおける GC 含量は各生物によって異なり、脊椎動物では３５～４５％と安定しているが、原核生物では２０～８０％と種によって非常に多様である。種による GC 含量の違いは、進化の過程で GC/AT 圧がかかったためと考えられるが、コード領域ではコドンのポジションごとに圧力の受け方は異なり、置換の多い第３ポジションで顕著に表れる（Muto and Osawa 1987)。厳密には GC 含量の違いだけでなく、ゲノムによって４種の塩基の組成には偏りがあり、そのような偏りが生じるメカニズムは鎖の対称性と関係している（Sueoka 1995; Lobry 1995, 1996)。

もし、塩基置換が DNA の二本鎖で対称的に起こると仮定すると、核ゲノムではリーディング鎖とラギング鎖、ミトコンドリアでは H 鎖と L 鎖において、Watoson-Crick 対の法則から転移確率は鎖対称的になる。例えば、A->G 転移が一方の鎖で生じる確率をP_{AG}とし、その相補鎖で生じる確率をQ_{AG}とすると、$P_{AG}=Q_{TC}$であり、同様に$Q_{AG}=P_{TC}$である。鎖対称性の仮定のもとでは$P_{AG}=Q_{AG}$であるから、$P_{AG}=P_{TC}$かつ$Q_{AG}=Q_{TC}$となり、６通りの全ての置換型でも同様である。

ここで、一本の鎖における A, T, G, C の頻度をそれぞれf_A，f_T，f_G，f_Cとすれば、A と T の頻度の変化は、

$$\Delta f_A = -(P_{AT}+P_{AG}+P_{AC})\cdot f_A + P_{TA}\cdot f_T + P_{GA}\cdot f_G + P_{CA}\cdot f_C$$

$$\Delta f_T = P_{AT}\cdot f_A - (P_{TA}+P_{TG}+P_{TC})\cdot f_T + P_{GT}\cdot f_G + P_{CT}\cdot f_C$$

であるが、Watoson-Crick 対的に塩基組成の偏りがなければ$f_A=f_T$，$f_G=f_C$であるから、$\Delta f_A=\Delta f_T$となり A と T の頻度変化量も等しくなる。同様に、$\Delta f_G=\Delta f_C$となるため、つねに塩基組成に偏りは生じない。多くの置換モデルは、このように塩基組成に偏りがなく、塩基置換も鎖対称的であると仮定している

が、実際には塩基組成に偏りがあり、それはDNAの複製、転写、組み換えが二本鎖間で異なるためであると考えられている（Wu and Maeda, 1987; Kunkel 1992; Waga and Stillman 1994; Beletskii and Bhagwat 1996; Francino and Ochman 1997; Freeman et al. 1998）。そして、非転写鎖ではC->Tデアミ化が起こりやすく（Beletskii and Bhagwat 1996）、鎖非対称的な複製の誤りが集団における遺伝的変異を増加させ、自然選択に対する適応性を高めているという仮説もある（Furusawa and Doi 1992, 1998）。

3.2 方法

もし塩基置換も鎖非対称に生じているとすれば、A と T、G と C の頻度の間に偏りが生じる。そして、複製単位が交換する点では、その偏りの傾向（Skew）が正反対に入れ替わるはずである。このような仮定のもとに、以下のようにそれぞれの Skew を定義し、

$$ATS = \frac{f_A - f_T}{f_A + f_T}$$

$$GCS = \frac{f_G - f_C}{f_G + f_C}$$

適当なウィンドウサイズで ATS や GCS を計算することによって、原核生物の複製の開始点が予測されている（Grigoriev 1998; McLean et al. 1998; Lopez et al. 2000）。McLean et al.（1998）は１２の原核生物で ATS と GCS を調べ、コドンの第三ポジションのみを用いると複製開始点の予測精度が高まることを示している。

本研究では、まず１５２のミトコンドリアの完全ゲノムを系統的に分類し、ゲノム全体における ATS と GCS を調べた。そして、ヒトとシロイヌナズナ、酵母のミトコンドリアでは、ウィンドウを作ってゲノム全体での ATS と GCS のパターンを調べた。次に、６６の原核生物の完全ゲノムについて、１０ｋｂのウィンドウサイズで、ATS と GCS さらにそれらを累積した値（cumulative skew）を計算し、グラフ化した。真核生物ではヒトの２１番、２２番、６番のコンティグについて、および線虫とシロイヌナズナの全染色体と酵母の６番染色体について同様に計算した。

さらに、３５９ミトコンドリア、１０５原核生物、そして９種の真核生物ゲノムにおける全コード領域を抽出し、各コード領域での ATS および GCS とコドンの各ポジション別に ATS と GCS を計算し、GC 含量との相関を調べた。

3.2 結果

ミトコンドリア

１５２のmtDNAは、哺乳類（１～４０）、鳥類（４１～５０）、爬虫類（５１～５６）、両生類（５１～５６）、魚類（５９～８０）、無脊椎動物（８１～１２０）、菌類（１２１～１２５）、植物（１２６～１３７）、および（原生生物１３８～１５２）である（表A.5）。有顎脊椎動物のmtDNAにおけるGC含量はみな４０%程度であり、遺伝暗号規則も同じである（Jukes and Osawa 1990; Osawa 1994）。その他の生物では、GC含量も多様であり、*Apis mellifera*（15.1%）が最も低く、*Balanoglossus camosus*（48.6%）が最も高い。遺伝暗号規則も多様化しており、表中のGはNCBIのデータベースによる分類番号であるが、Jukes and Osawa(1993)に基づいている。

無脊椎動物や植物を除く多くの種では、ATSとGCSの値が明らかに0から外れており、ATSとGCSの値は完全に異なる。これらのパターンはATS > 0かつGCS < 0、もしくはATS < 0かつGCS > 0であり、ウィンドウを作ってATSとGCSの値を調べればそれぞれ単調に増加または減少する。一方の鎖でATS, GCS > 0であるとき、もう一方の鎖ではATS、GCS < 0であるため、２つのパターンはH鎖とL鎖の違いだけであり、基本的には同じパターンであると考えられる。ATSとGCSで明らかに偏りのパターンが生じるのは、鎖によって複製時の誤りのパターンが異なるためであり、転写時の誤りはほとんど関係していない（Francino and Ochman 1997）。

植物の mtDNA のゲノムサイズは動物のものに比べて１０から２０倍と大きく、３つの被子植物では、|ATS|, |GCS| ≦ 0.01であり全体の偏りがほとんどない。植物の mtDNA の複製様式については明らかではないが、シロイヌナズナではATSとGCSのパターンがヒトのmtDNAのパターンのように単調ではなく複雑であるため、植物のmtDNAでは複製の開始点が複数あり全体で偏りを打ち消しあっていると考えられる。また、遺伝暗号の規則も核ゲノムと同じ標準規則であり、ATSやGCSのパターンも動物のmtDNAよりは核ゲノムのパターンに似ているため、複製様式も核ゲノムに似ている可能性が高い。酵母（*Saccharomices cerevisiae*）のmtDNAは特異的で、ATSとGCSのパターンは非常に複雑である。複製開始点が３つと複製開始点に似た領域が４つあり、複製の方向が２方向である可能性が示されているが（Lecrenier and Foury 2000）、全体のATSとGCSの値は打ち消しあっておらず正のままである。このことは、複製の開始点の数だけでなく、様式や方向性に関する情報が重要であることを示唆する。複製の開始点が複数あったとしても、各々の点で鎖の複製様式が入れ替わらなければ、偏りのパターンも複製開始点で反転せずに傾向を保ち続ける。酵母のmtDNAの複製システムについては、どちらかといえば動物のmtDNAに似ており、植物の

mtDNA とは異なる。

また、359 種のミトコンドリアゲノムにおける CDS 全体の GC 含量と各コドンポジション別 GC 含量との相関（平均値）では、第 1 ポジションの GC 含量は高く、第 3 ポジションは全体に GC 含量が低くて傾きが大きい。第 2 ポジションは、中間で傾きが小さい（図 3.1）。CDS の GC 含量と各コドンポジション別 ATS と GCS の相関について、ATS は第 2 ポジション以外正の相関があり、GCS は全て負の相関である。また、第 2 ポジションはいずれも負の領域に分布が多く、A<T かつ G<C の傾向が強い（図 3.2）。

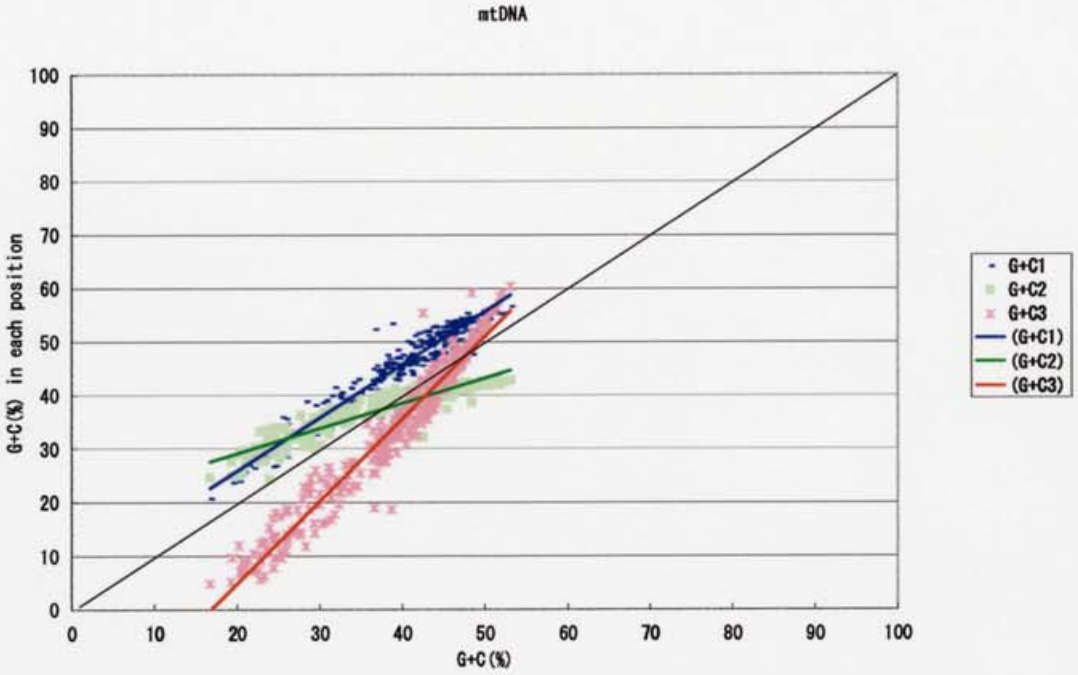


図 3.1 359 種のミトコンドリアゲノムにおける CDS 全体の GC 含量と各コドンポジション別 GC 含量との相関（平均値）。第 1 ポジションの GC 含量は高く、第 3 ポジションは、全体に GC 含量が低くて傾きが大きい。第 2 ポジションは、中間で傾きが小さい。

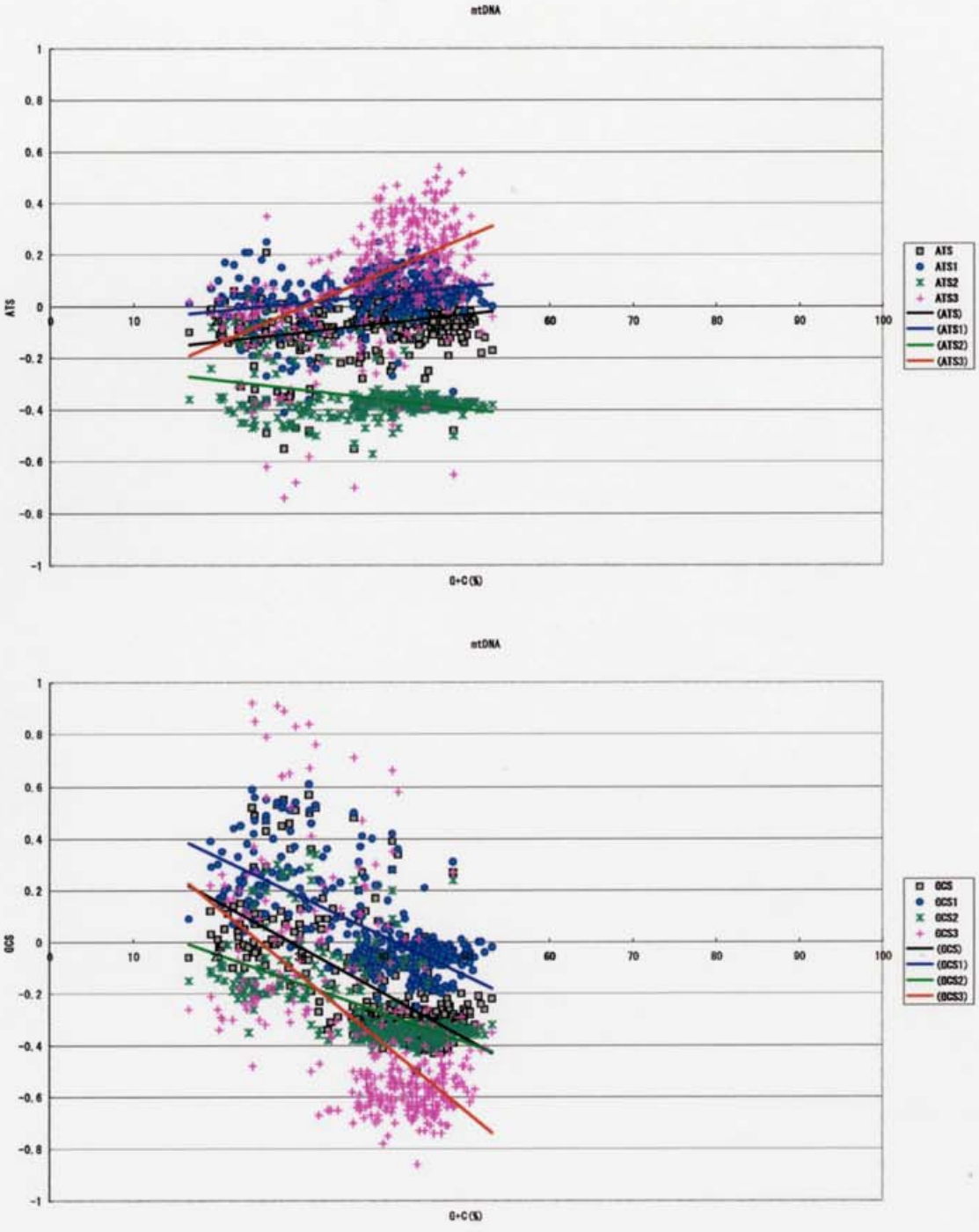


図 3.2　359 種のミトコンドリアゲノムにおける CDS の GC 含量と各コドンポジション別 ATS（上図）と GCS(下図)の相関。ATS については、第 2 ポジション以外は正の傾きであり、GCS については全て負の傾きである。また、第 2 ポジションはいずれも負の領域に分布が多く、A<T かつ G<C の傾向が強い。

原核生物

原核生物のゲノムは、ほとんど全体がコード領域であるが、そのゲノムサイズや塩基組成は多様である。高熱環境に生息する原核生物では、熱に対する結合力の安定性から高 GC のコドンが多く GC 含量が高くなり、生息環境や系統関係とも相関している。

原核生物ではいずれも、全体の ATS と GCS の値がほとんど 0 である。しかしながら、ウィンドウを作って全体のパターンを調べると、鎖によって ATS または GCS が非常に強く Skew を持ち、それらの傾向が複製単位で入れ替わることによって、全体で打ち消しあっていることがわかる(McLean et al. 1998)(図 3.3)。多くの真正細菌では、ほぼ一定の強い ATS または GCS がゲノムのちょうど半分ずつで入れ替わるため、容易に複製開始点を予測することができる（Grigoriev 1998)。しかしながら、古細菌では、ATS と GCS のパターンも複雑であり、複製の開始点の予測はきわめて難しい。古細菌の複製様式については不明であるが、細菌よりは真核生物の様式に似ていると言われており、Skew のパターンが真核生物の核ゲノムに似ていることはこの説を支持する。

また、明確な Skew のパターンを持つ真正細菌において、Proteobacteria の α, β, γ-subdivision , Actinobacteria, Chlamydiales, Spirochaetales では ATS と GCS の符号は反転しているため、Skew のパターンは転移によって生じた可能性が高い（付録 B 参照)。大腸菌には突然変異のホット・スポットである Chi 配列と呼ばれる 5'-GCTGGTGG-3'配列があり（Horiuchi 1995)、この配列の分布が GCS のパターンをもたらしているようである。しかし、Bacillus グループでは ATS と GCS の符号が同じであるため、A と G が同時に増加/減少している。つまり、リーディング鎖とラギング鎖において、一方にプリンが多くもう一方にピリミジンが多い非対称性であるため、塩基置換も転換型が多いと推定できる。

コード領域における GC 含量、GCS、ATS をポジション別に計算した値の散布図も種によって多様である。大腸菌や枯草菌では、全体の GC 含量は異なるが、いずれも第 1、第 3、第 2 ポジションの順である。大腸菌ゲノムの第 3 ポジションは特に傾きが強い(図 3.4)。ATS は全体の GC 含量と負の相関になっているが、GCS は第三ポジション以外では正の相関になっている。更に詳しくみると、ATS では、第 1、第 2、第 3 の順であり、GCS では、第 1、第 3、第 2 の順である。つまり、第 1 では A や G が多く、第 2 では C、第 3 では T が多い（図 3.5、図 3.7)。

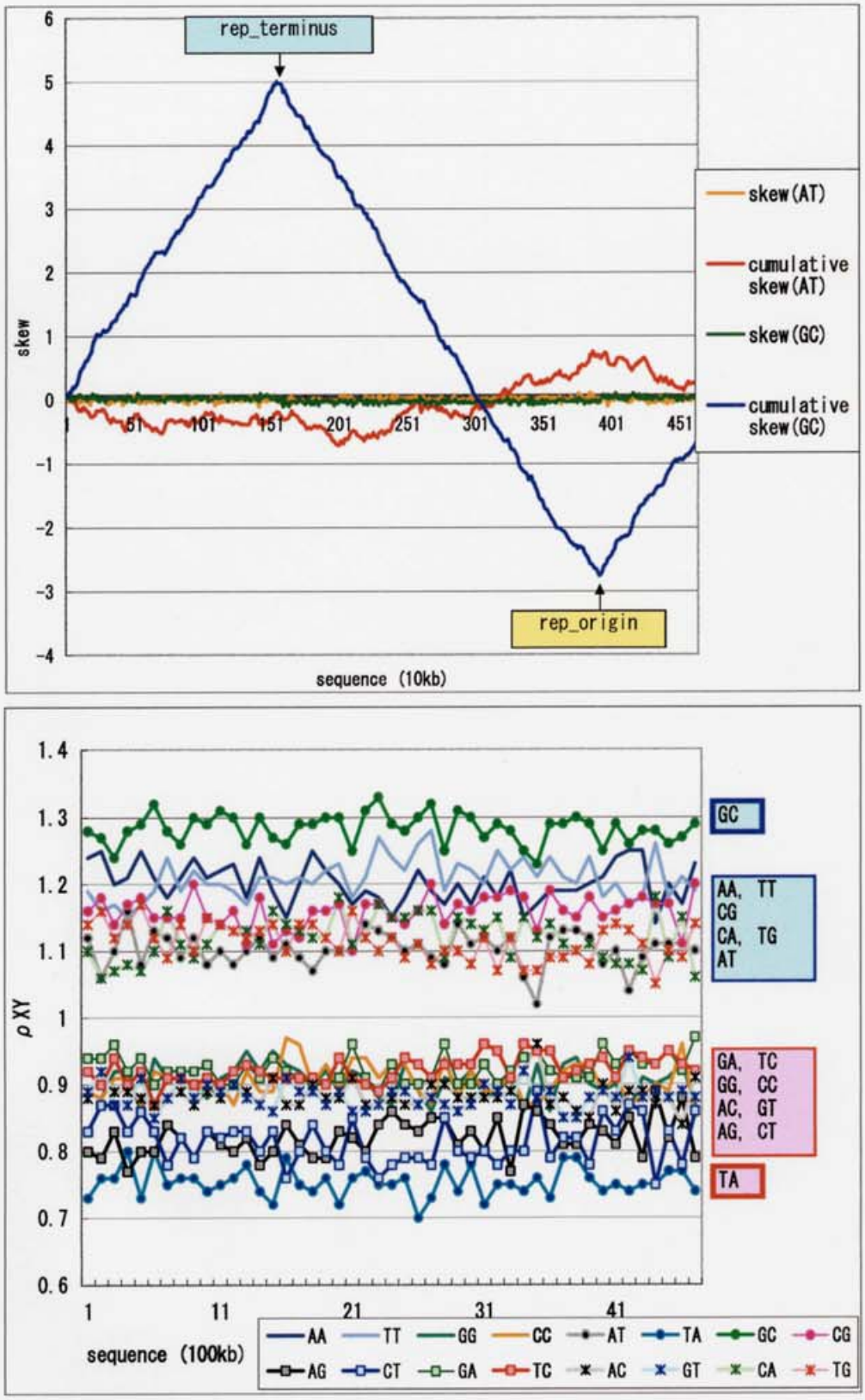


図 3.3 大腸菌ゲノムにおける Skew(上図)と二連塩基の頻度/期待値（下図)。

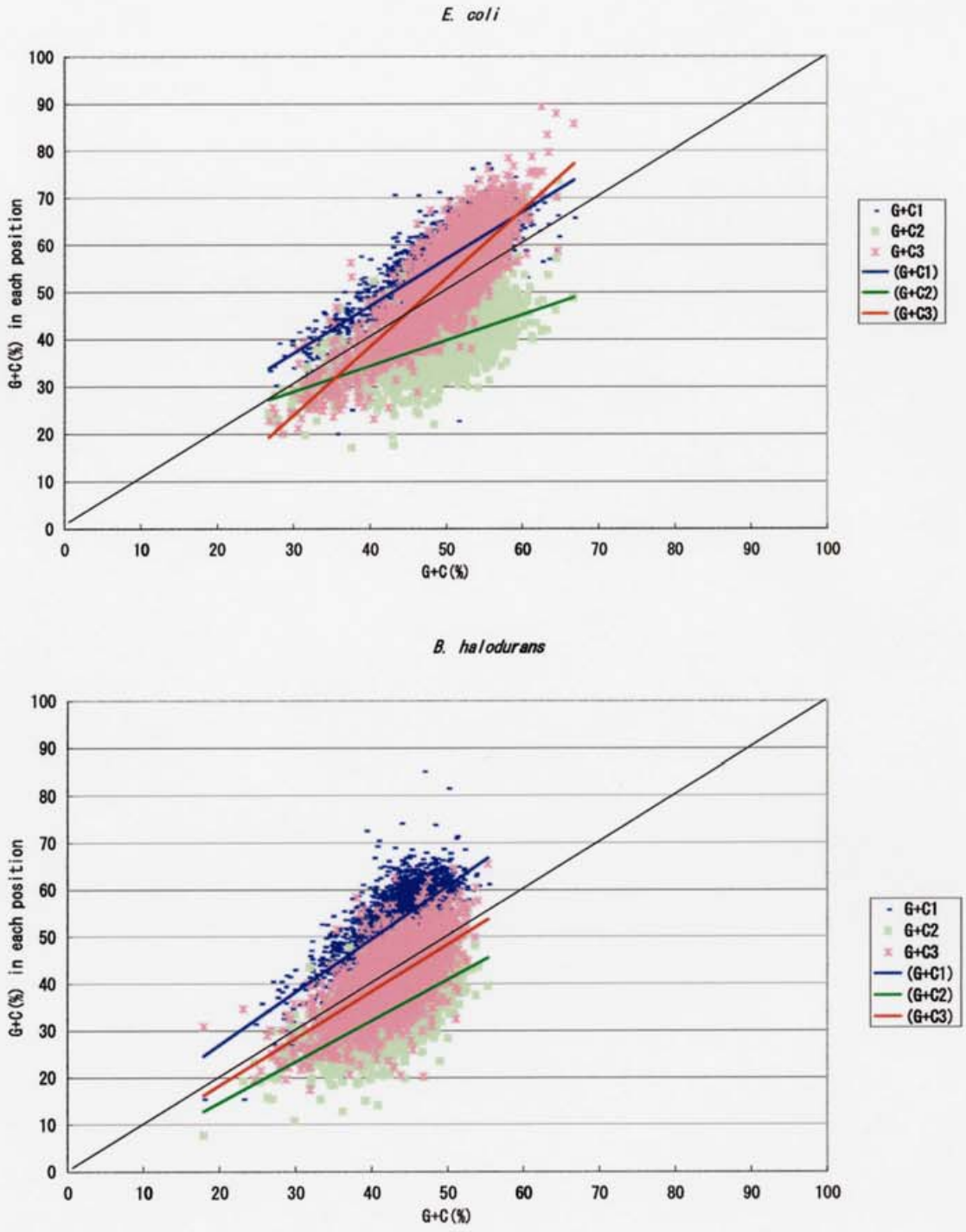


図 3.4 大腸菌ゲノム（上図）と枯草菌ゲノム（下図）における CDS の GC 含量と各コドンポジション別 GC 含量との相関。全体の GC 含量は異なるが、いずれも第 1、第 3、第 2 ポジションの順である。大腸菌ゲノムの第 3 ポジションは特に傾きが強い。

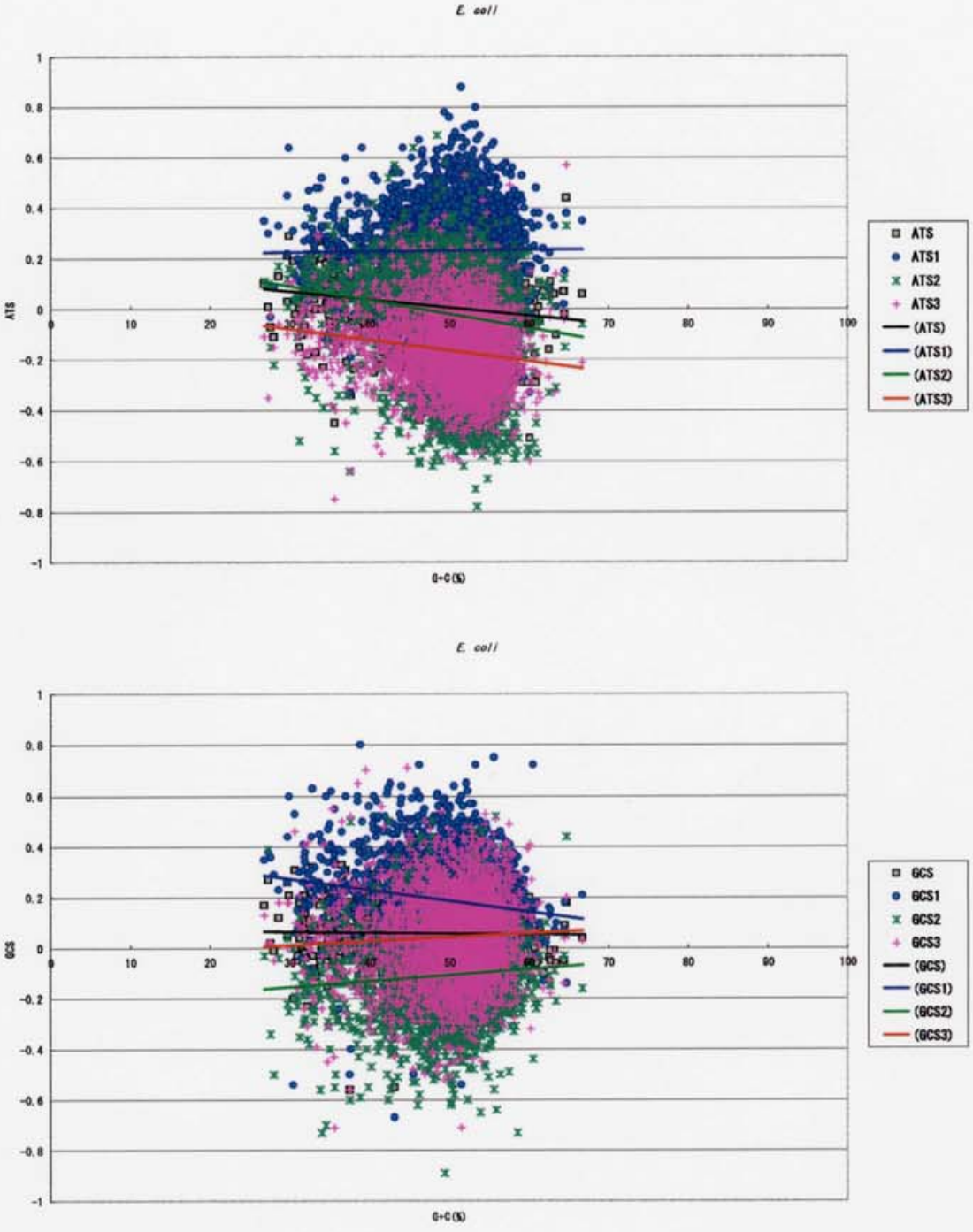


図 3.5 大腸菌ゲノムにおける CDS の GC 含量と各コドンポジション別 ATS（上図）と GCS（下図）の相関。ATS では、第 1、第 2、第 3 の順であり、GCS では、第 1、第 3、第 2 の順である。つまり、第 1 では A や G が多く、第 2 では C、第 3 では T が多い。

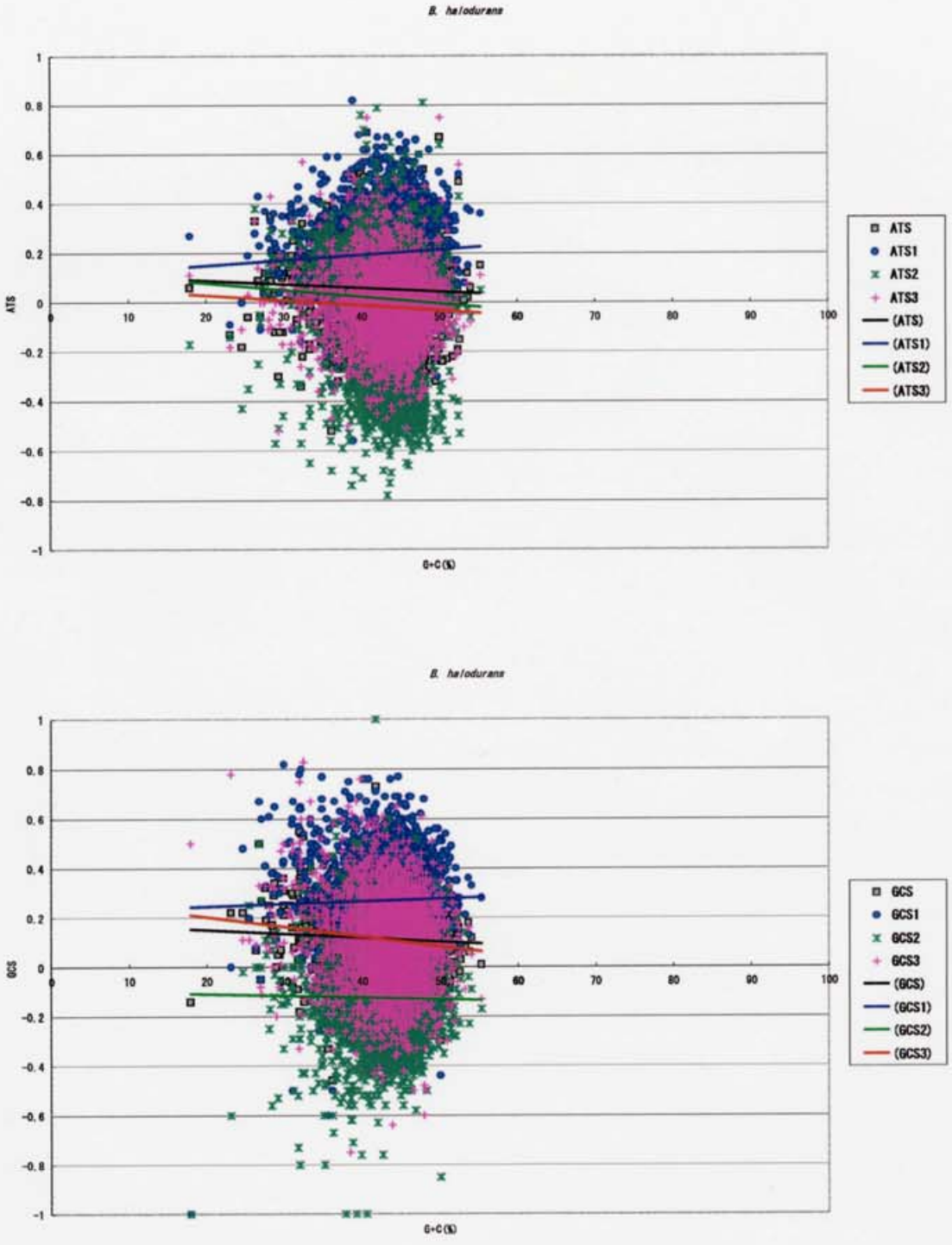


図 3.6 枯草菌ゲノムにおける CDS の GC 含量と各コドンポジション別 ATS（上図）と GCS（下図）の相関。大腸菌と同じく、ATS では、第 1、第 2、第 3 の順であり、GCS では、第 1、第 3、第 2 の順である。

ヒトゲノム

国際的なヒトゲノムプロジェクトの成果が発表され（nature 2001）、本年中にはヒトゲノムの完全配列が決定される予定である。既に配列が決定された代表的なコンティグとして６番染色体の HLA 領域４．４Mb(図 3.7)、２１番染色体の長腕２８．５Mb(図 3.8)、２２番染色体の長腕２３Mb(図 3.9)のパターンを調べた。HLA 領域は、ゲノム中で最も遺伝子密度が高く、CpG アイランドの分布や、遺伝子の機能的な解析や多型など、最も研究が進んでいる。

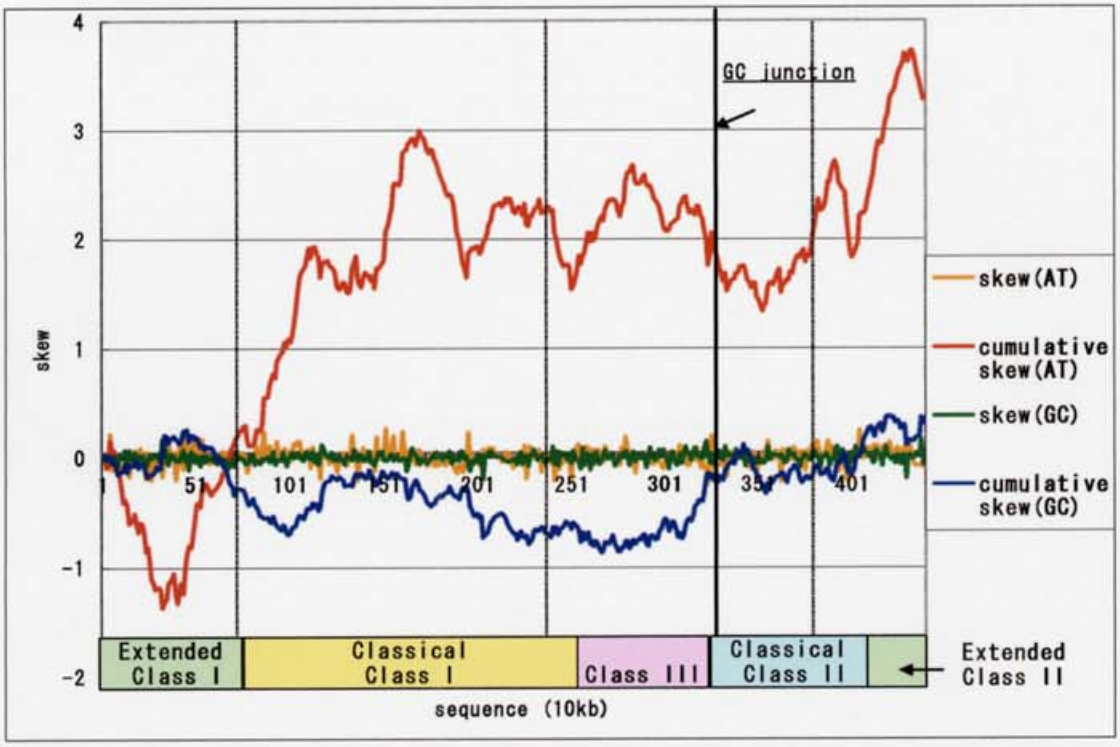
ヒトをはじめ高等生物のゲノムは、大きくアイソコアによって GC 含量が異なり、染色体のバンド構造や複製時期の違いと関係している（Bernardi et al. 1985）。恒温動物と変温動物とでは異なるが、ヒトゲノムでは、低 GC（GC-poor）のアイソコア L_1 と L_2 が全体の６２％を占め、高 GC（GC-rich）の H_1 と H_2 が２２％と９％であり、最も高 GC な H_3 は３～４％である（Bernardi 1995）。染色体のバンド構造は、ギムザ染色による違いから G バンドと R バンドに分けられ、さらに R バンドの一部は T バンドに分けられる。G バンドは L_1 と L_2 から、R バンドは H_1 と L_1、L_2 から、T バンドは主に H_2 と H_3 から構成されている（Saccone et al. 1993）。T バンドでは GC 含量が高く、細胞周期の早い時期に複製が行われるが、G バンドでは GC 含量が低く、複製の行われる時期も遅い（Bernardi et al. 1985）。そして、一般に遺伝子の GC 含量はその遺伝子が位置しているバンドもしくはアイソコアの GC 含量と相関している。

HLA 領域でのバンド境界の一つとして、クラス III とクラス II の遺伝子クラスター間の境界があり、そこでは構造的な GC 含量の変化と複製時期の変化がみられる（Ikemura et al. 1990; Fukagawa et al. 1995）。この境界は、コンティグのセントロメア側（右）から 1 Mb ほどにあり、この付近では正の累積値を持つ ATS が急に ATS < 0 になり、0 に近い負の累積値を持つ GCS が急に GCS > 0 になる。テロメア方向にいくと、ATS や GCS の累積値でいくつもの極大・極小があらわれ、クラス I の遺伝子クラスター領域の境界付近においても両側で極値になっている。このような極値では、複製時期あるいは複製単位の変化が起きていると推測される。他のコンティグについても共通に言えることであるが、ATS のほうが GCS よりも変動のしかたが大きいことは、AT 圧が強くかかったためと考えられる。

２１番のコンティグは、GC 含量が領域によって大きく異なり、右側 1/3 ほどの高 GC 領域以外は、極端に遺伝子密度が小さい遺伝子砂漠である。２２番のコンティグに比べ Skew の度合いは大きく、局所的な変動はあるものの傾向は ATS > 0 と GCS < 0 の一貫性が強い。つまり、複製単位は複数あるはずであるが、二

本鎖の一方ではリーディング鎖となる領域が多く、もう一方ではラギング鎖となる領域が多くなっていると考えられる。２２番では、ATS と GCS の累積値のパターンが複雑に何度もクロスしているため、原核生物ゲノムのように、一本の鎖でリーディング鎖となる領域とラギング鎖となる領域が複雑に何度も入れ替わり、均等的になっていると考えられる。

コード領域における GC 含量、GCS、ATS をポジション別に計算した値の散布図からは(図 3.10)、まず第３ポジションでは GC 含量が高く、分散も大きくて、全体の GC 含量を左右していることがわかる。そして、第２ポジションは全体に GC 含量が低めであり、GCS と ATS はいずれも全体の GC 含量と負の相関になっているが、第二ポジションの GCS は正の相関を持つ染色体が多い。Skew についてまとめると、第１ポジションは A と G のプリンが多く、第２ポジションは C が多くなり、第３ポジションは T が多くなっている(図 3.11)。


GC junction
skew
sequence (10kb)
skew(AT)
cumulative skew(AT)
skew(GC)
cumulative skew(GC)
Extended Class I
Classical Class I
Class III
Classical Class II
Extended Class II

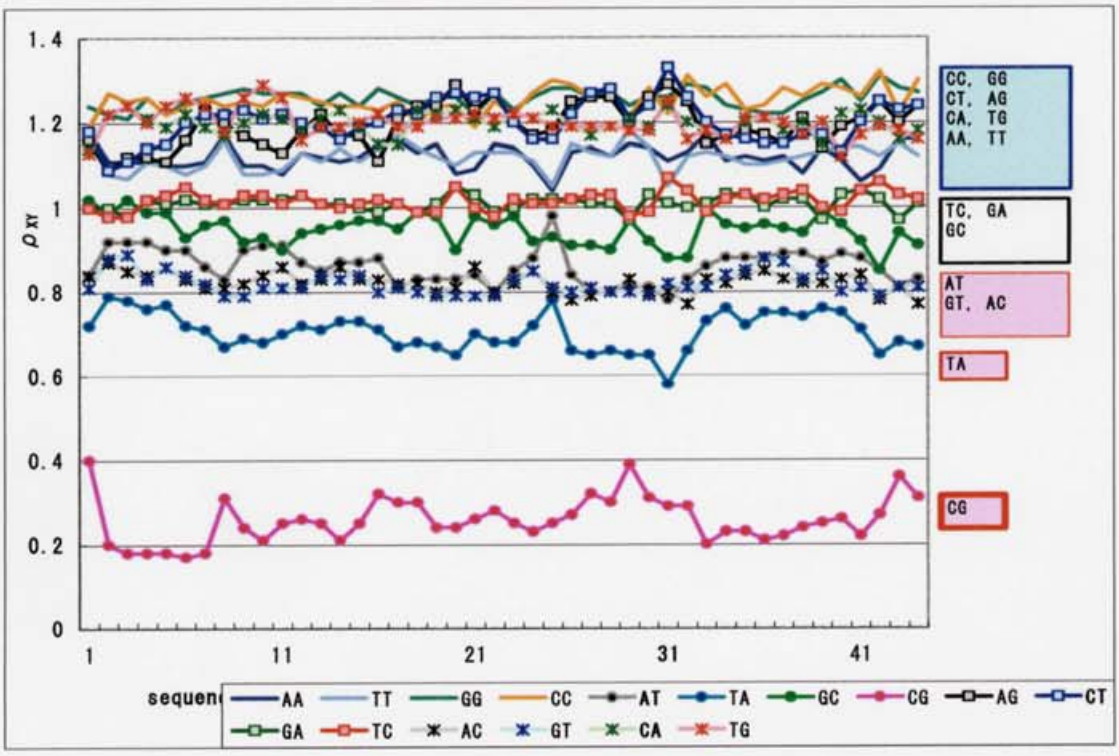
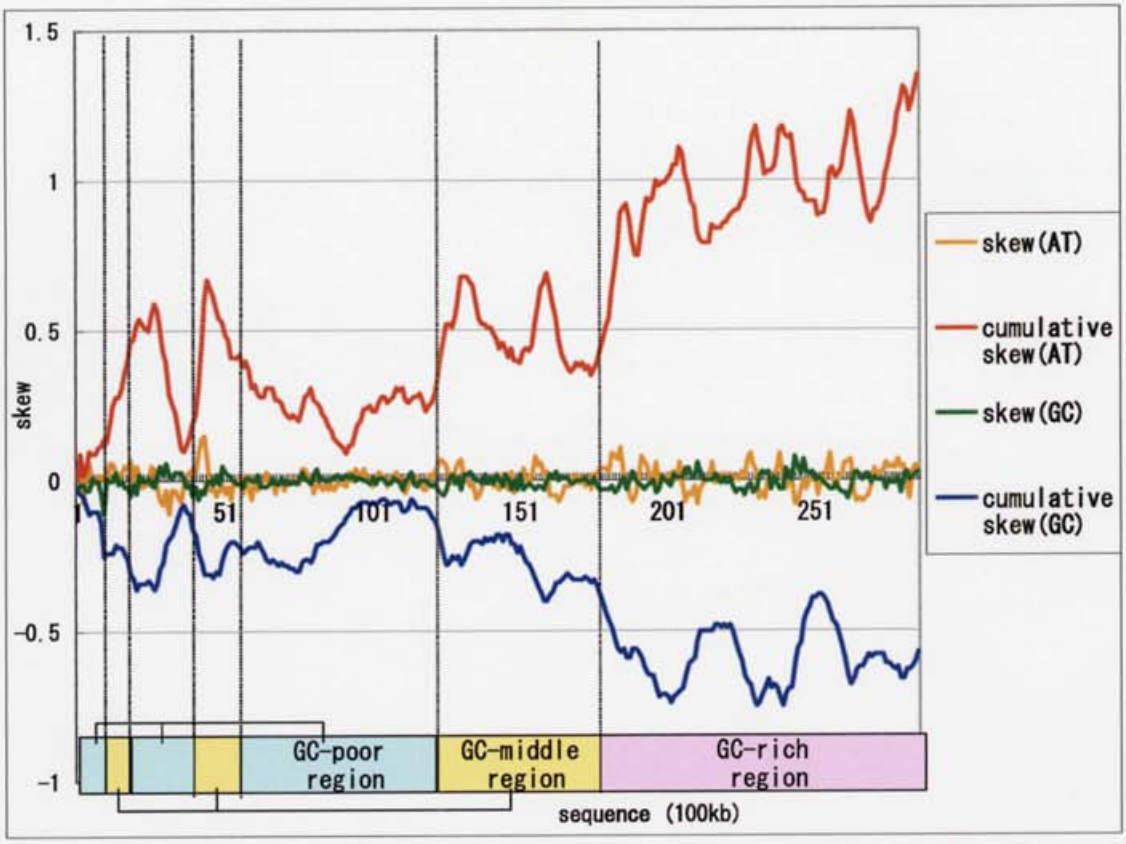

ρ xy
CC, GG
CT, AG
CA, TG
AA, TT
TC, GA
GC
AT
GT, AC
TA
CG
AA
TT
GG
CC
AT
TA
GC
CG
AG
CT
GA
TC
AC
GT
CA
TG

図 3.7 ヒト 6 番染色体（HLA 領域）における Skew（上図）と二連塩基の頻度/期待値（下図）。HLA 領域でのバンド境界の一つとして、クラス III とクラス II の遺伝子クラスター間の境界では構造的な GC 含量の変化と複製時期の変化がみられる。この境界は、コンティグのセントロメア側（右）から 1 Mb ほどにあり、この付近では正の累積値を持つ ATS が急に ATS < 0 になり、0 に近い負の累積値を持つ GCS が急に GCS > 0 になる。テロメア方向にいくと、ATS や GCS の累積値でいくつもの極大・極小があらわれ、クラス I の遺伝子クラスター領域の境界付近においても両側で極値になっている。このような極値では、複製時期あるいは複製に関するマクロな構造の変化が起きている可能性が高い。


1.5
1
0.5
0
-0.5
-1
skew
1
51
101
151
201
251
skew(AT)
cumulative skew(AT)
skew(GC)
cumulative skew(GC)
GC-poor region
GC-middle region
GC-rich region
sequence (100kb)

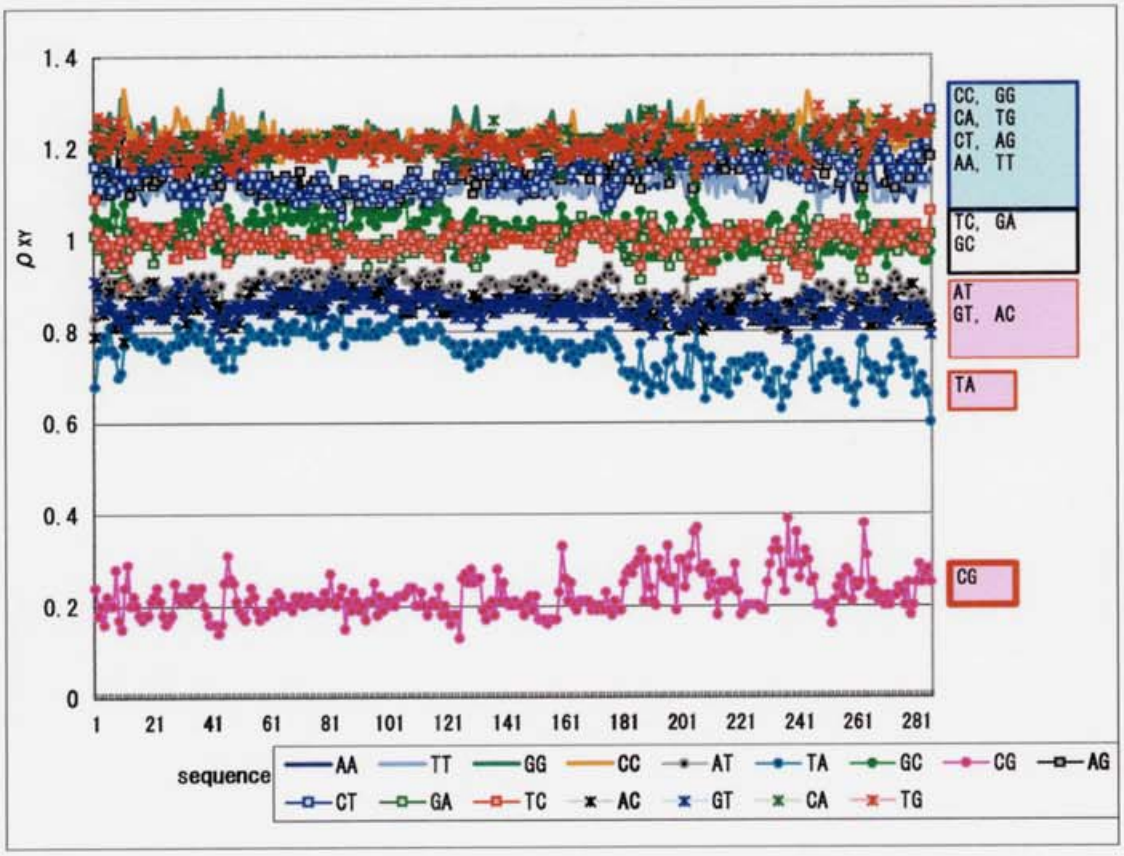
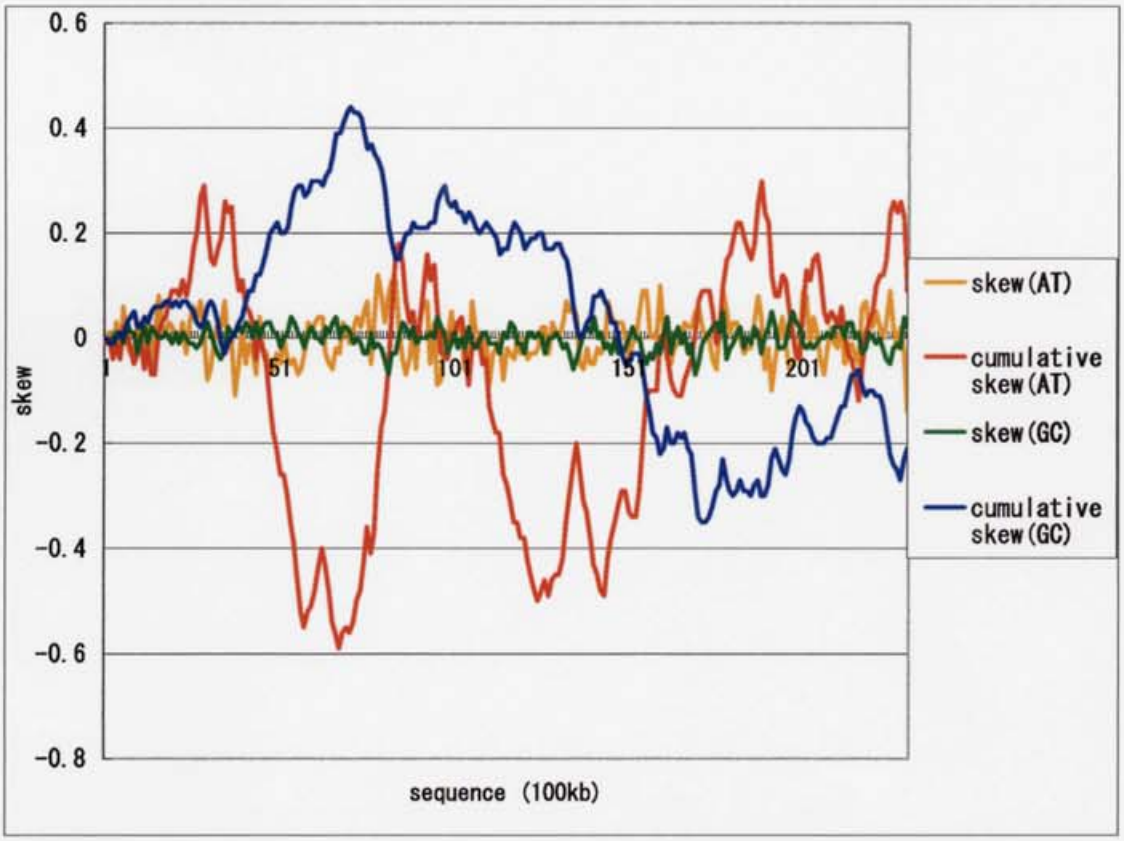

1.4
1.2
1
0.8
0.6
0.4
0.2
0
ρ XY
1
21
41
61
81
101
121
141
161
181
201
221
241
261
281
CC, GG
CA, TG
CT, AG
AA, TT
TC, GA
GC
AT
GT, AC
TA
CG
sequence
AA
TT
GG
CC
AT
TA
GC
CG
AG
CT
GA
TC
AC
GT
CA
TG

図 3.8 ヒト 21 番染色体における Skew(上図)と二連塩基の頻度/期待値（下図)。左端は長腕におけるセントロメア側であり、右端はテロメア側である。左側 2/3 ほどは、遺伝子砂漠と呼ばれる遺伝子密度のきわめて少ない領域である。ここでは、二連塩基に頻度の変動がほとんどみられない。


0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
skew
51
101
151
201
sequence (100kb)
skew(AT)
cumulative skew(AT)
skew(GC)
cumulative skew(GC)

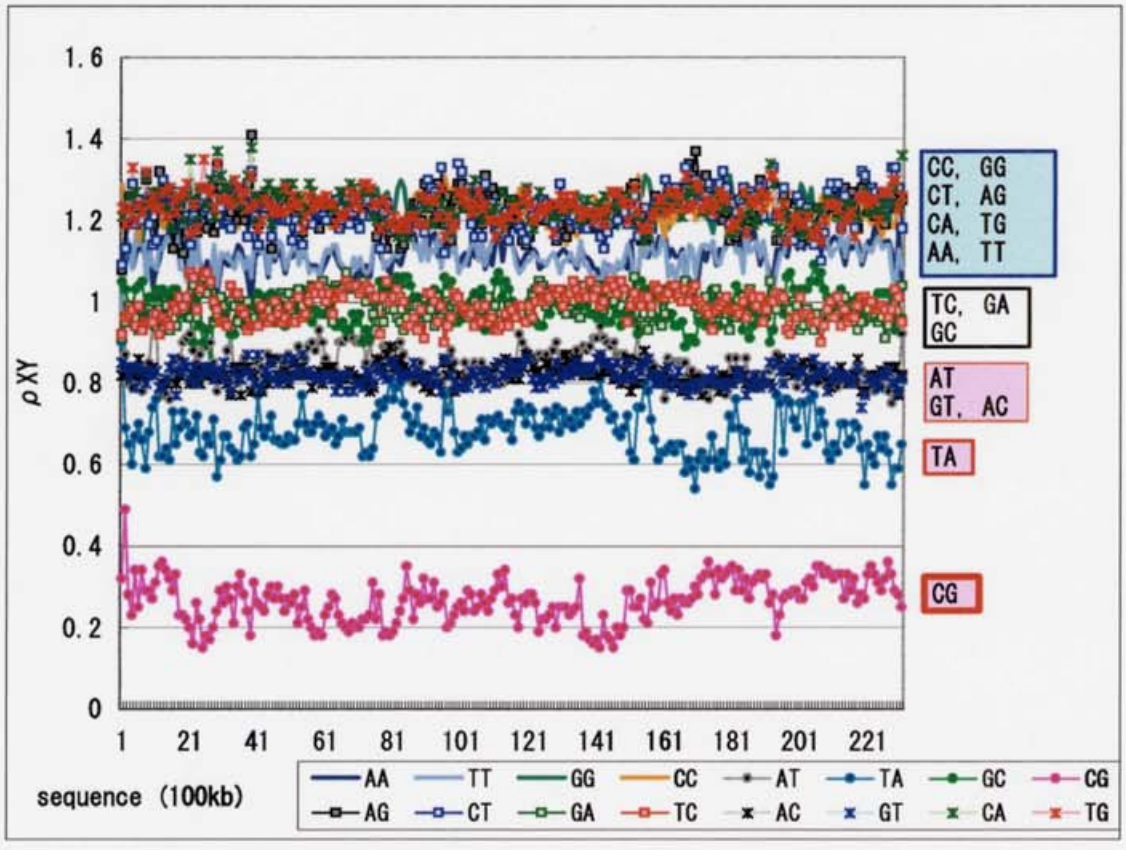

1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
ρXY
1
21
41
61
81
101
121
141
161
181
201
221
sequence (100kb)
CC, GG
CT, AG
CA, TG
AA, TT
TC, GA
GC
AT
GT, AC
TA
CG
AA
TT
GG
CC
AT
TA
GC
CG
AG
CT
GA
TC
AC
GT
CA
TG

図 3.9 ヒト 22 番染色体における Skew(上図)と二連塩基の頻度/期待値（下図)。Skew のパターンに 21 番染色体との違いがみられる。二連塩基の頻度傾向は同じであるが、２１番に比べ遺伝子密度が高いせいか小刻みな波が激しい。。

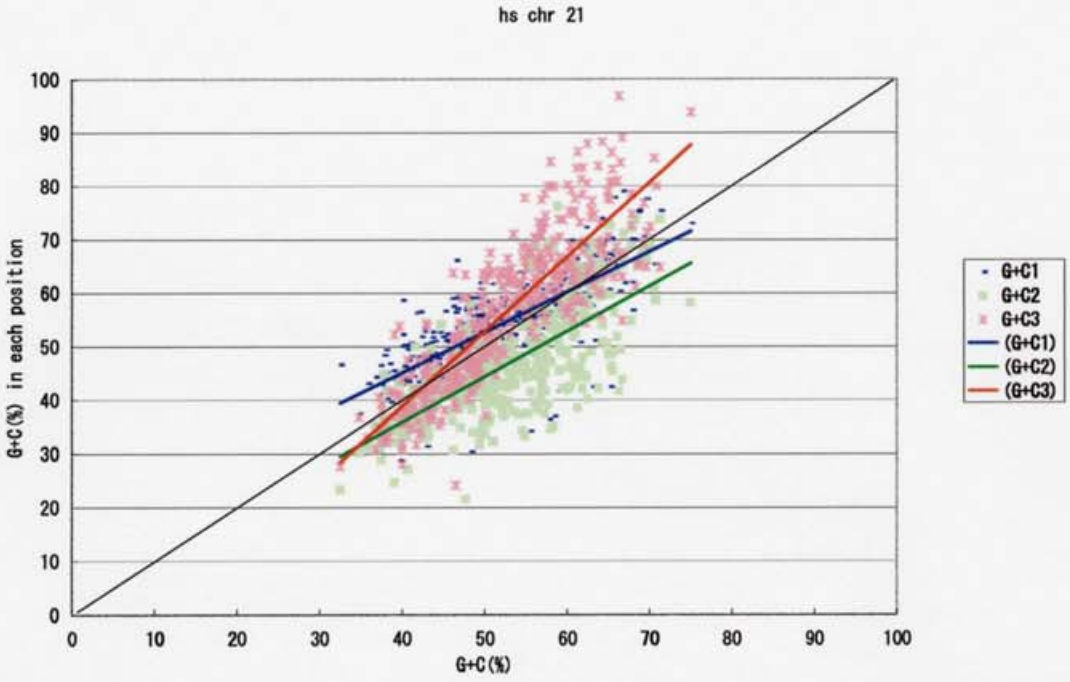


図 3.10 ヒト 21 番染色体における CDS の GC 含量と各コドンポジション別 GC 含量との相関。第 3 では傾きが大きく GC 含量も高めであり、第 2 では低めである。

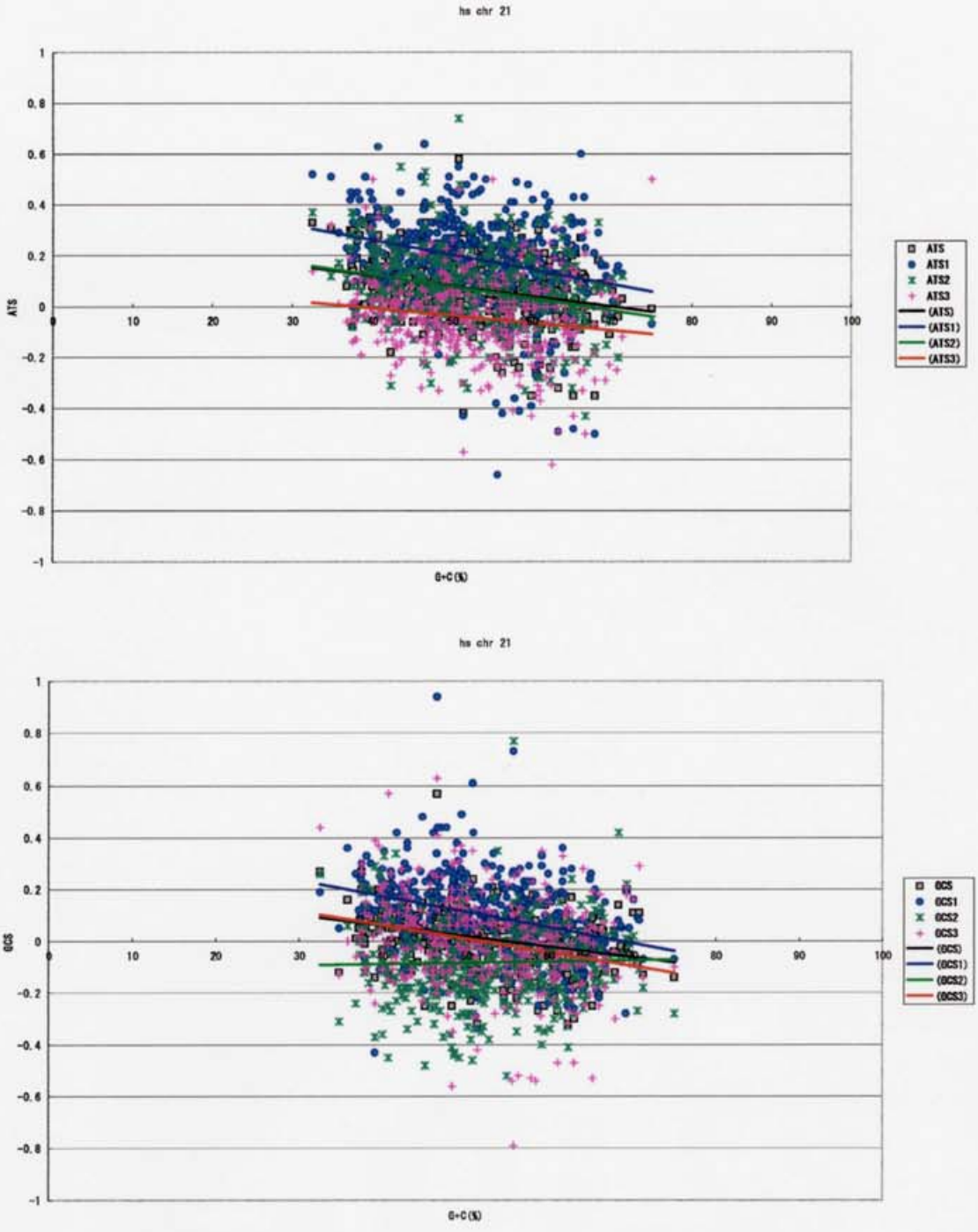


図 3.11 ヒト 21 番染色体における CDS の GC 含量と各コドンポジション別 ATS（上図）と GCS（下図）の相関。第 2 ポジションの GCS 以外は、いずれも負の相関である。大腸菌や枯草菌と同様、ATS では、第 1、第 2、第 3 の順であり、GCS では、第 1、第 3、第 2 の順である。

その他の真核生物

ゲノム配列全体のなかで、シロイヌナズナの完全ゲノムはデータが非常に高精度である。染色体全体に分布が調べられているGC含量を左右するトランスポゾンの分布と、GCS のパターンとの間に強い相関性がある。各染色体で共通して、セントロメアから少し離れたところでATSやGCSのパターンが強くなり、テロメアに近づくにつれ徐々にゆるくなっている(図3.12、付録B)。線虫のゲノム配列のデータはあまりにもギャップが多かったため、Skewのパターンは信頼性が低い（付録B）。酵母の6番染色体では、複製の開始点が実験的に調べられているが、Skewのパターンの極値と非常に相関性が高い。

コード領域におけるGC含量、GCS、ATSをポジション別に計算した値の散布図では、ヒトとは異なる特徴がある（図3.12～図3.15)。まず、シロイヌナズナや酵母では第一ポジションのGC含量のほうが、第3ポジションのGC含量よりも高い（図3.12)。更に、*P. falciparum*や*S. pombe*では、第三ポジションのGC含量が最も低くなる(図3.14)。

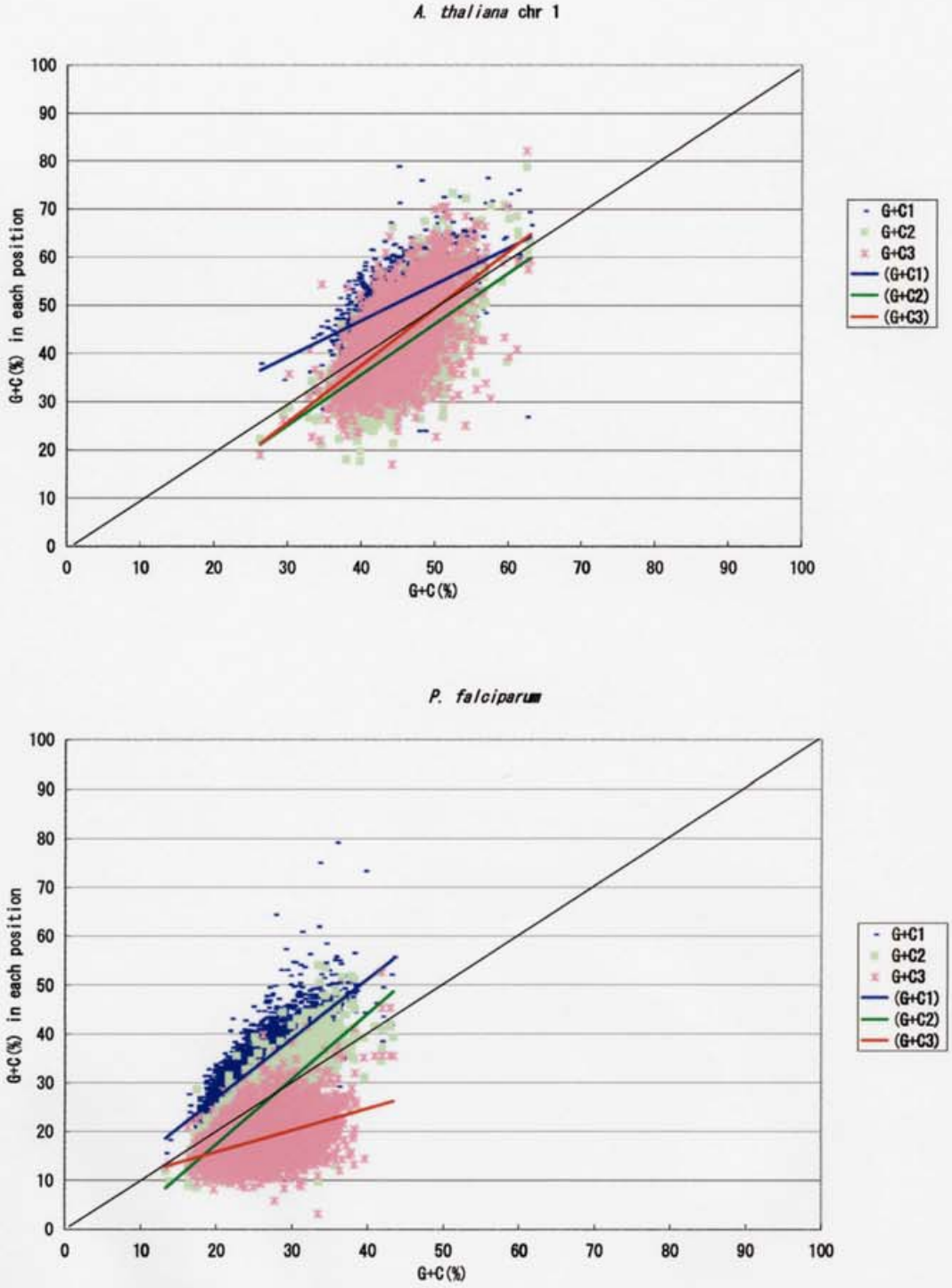


図 3.12 *A.thaliana* の第 1 番染色体（上図）と *P.falciparum* ゲノム（下図）における CDS の GC 含量と各コドンポジション別 GC 含量の相関。*A.thaliana* では、第 1、第 3、第 2 の順であり、*P.falciparum* は全体の GC 含量が低く、第 3 ポジションのの GC 含量も低く傾きも小さい。

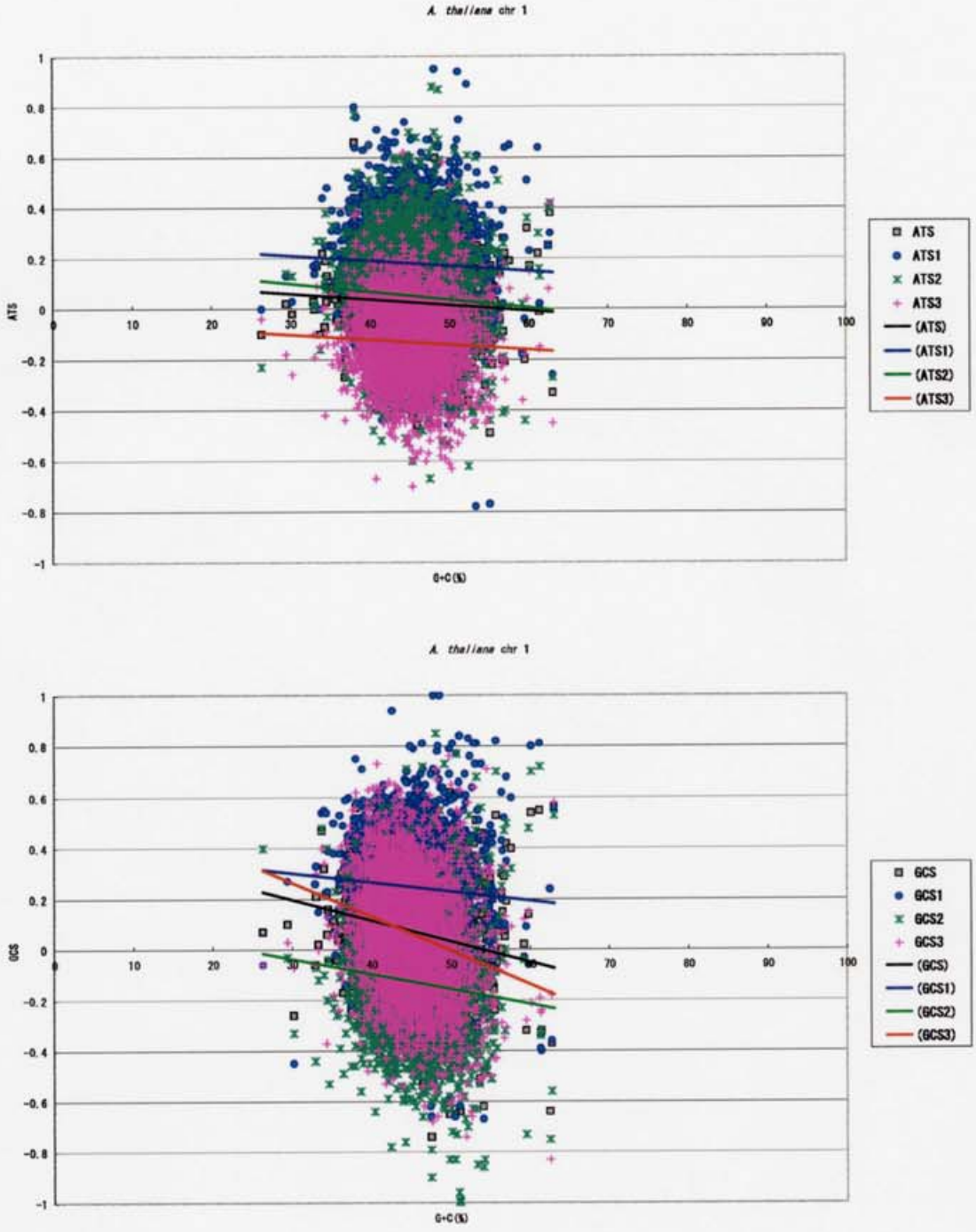


図 3.13 *A.thaliana* の第 1 染色体における CDS の GC 含量と各コドンポジション別 ATS（上図）と GCS（下図）の相関。ATS では、第 1、第 2、第 3 の順であり、GCS では第 1、第 3、第 2 の順である。全て負の相関である。

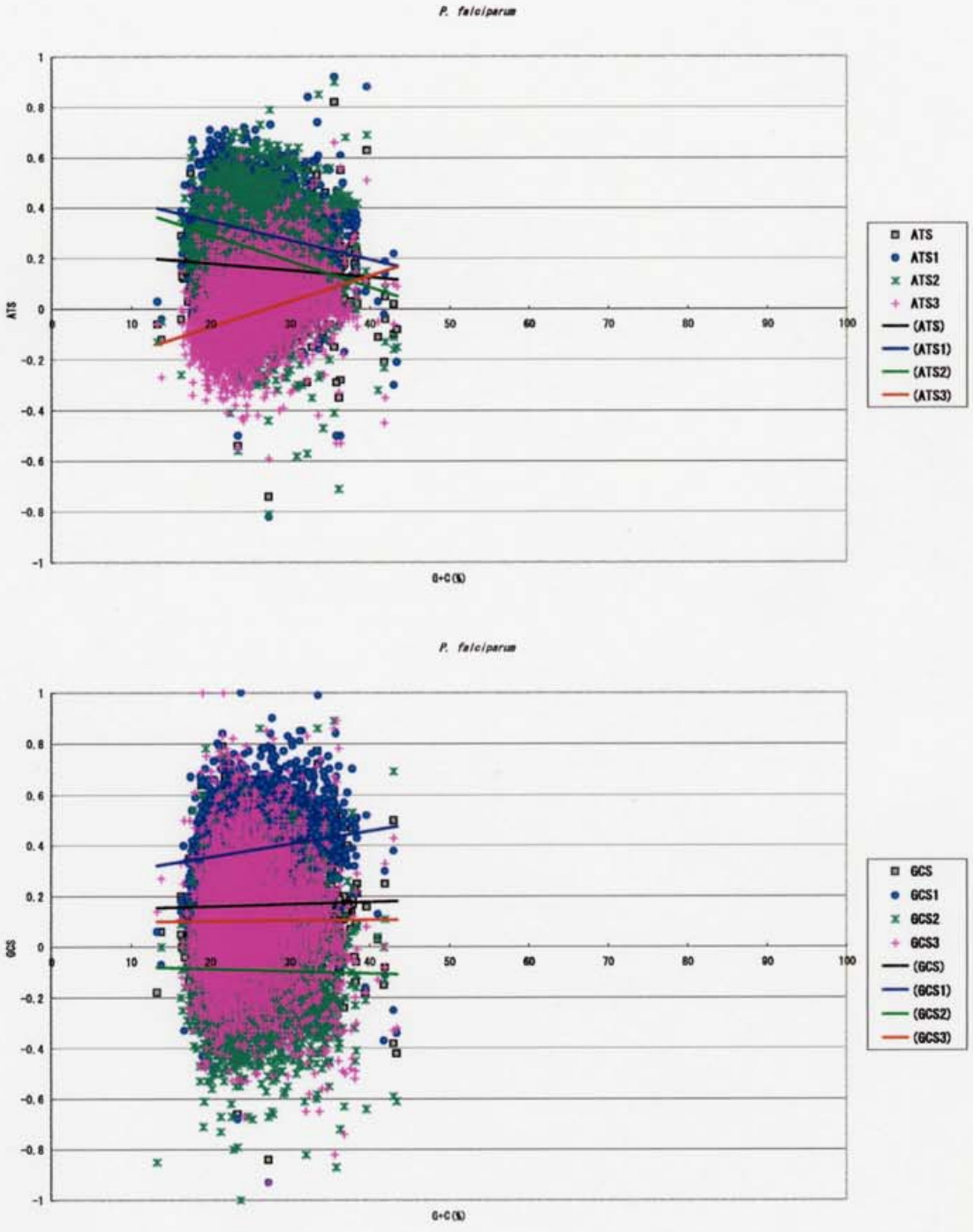


図 3.14 *P.falciparum* における CDS の GC 含量と各コドンポジション別 ATS（上図）と GCS（下図）の相関。ATS では、第 1、第 2、第 3 の順であり、GCS では、第 1、第 3、第 2 の順である。

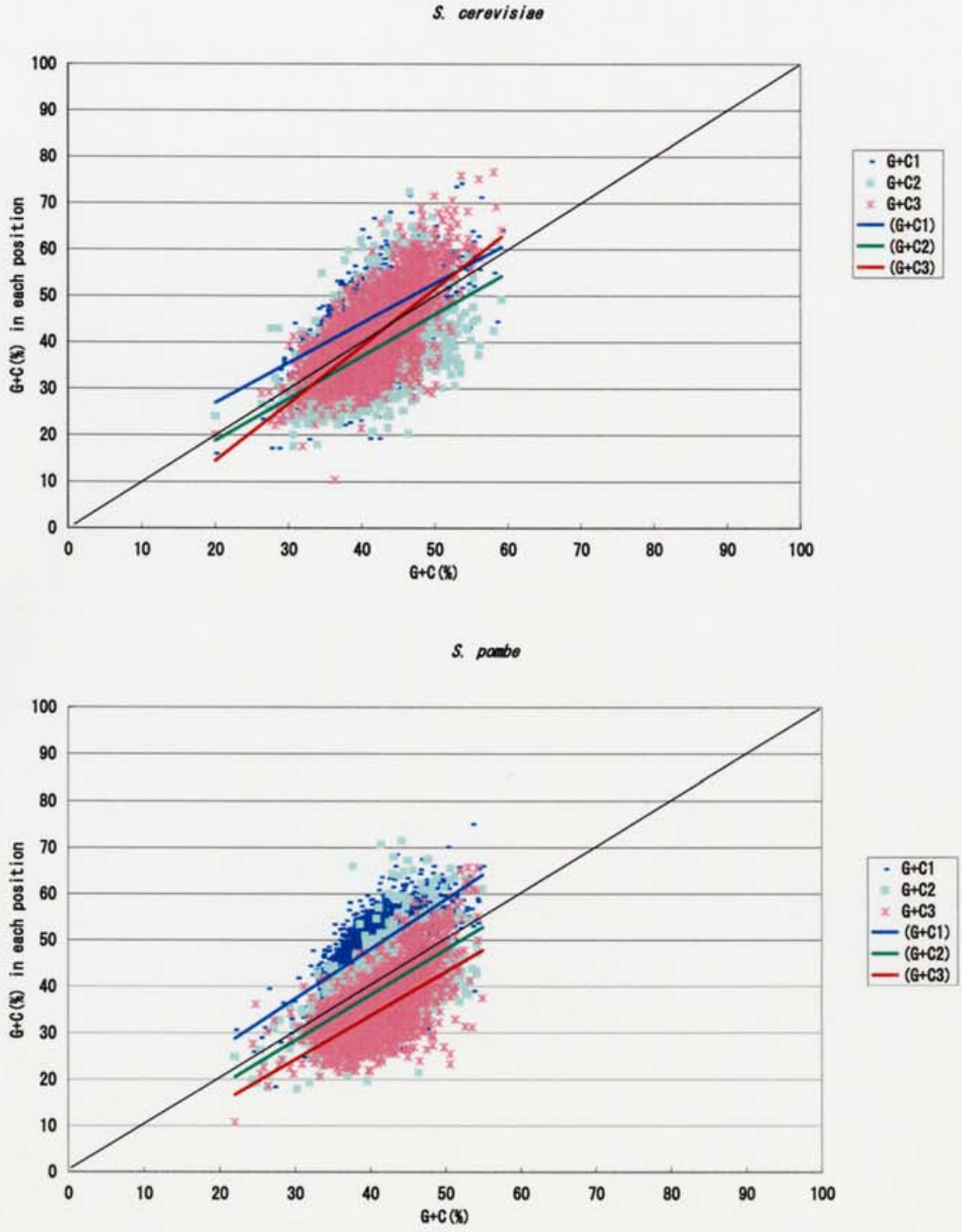


図 3.15 *S.cerevisiae*（上図）と *S.pombe*（下図）における CDS の GC 含量と各コドンポジション別 GC 含量の相関。*S.cerevisiae* では、第 1、第 3、第 2 の順であり、*S.pombe* では第 1、第 2、第 3 の順である。

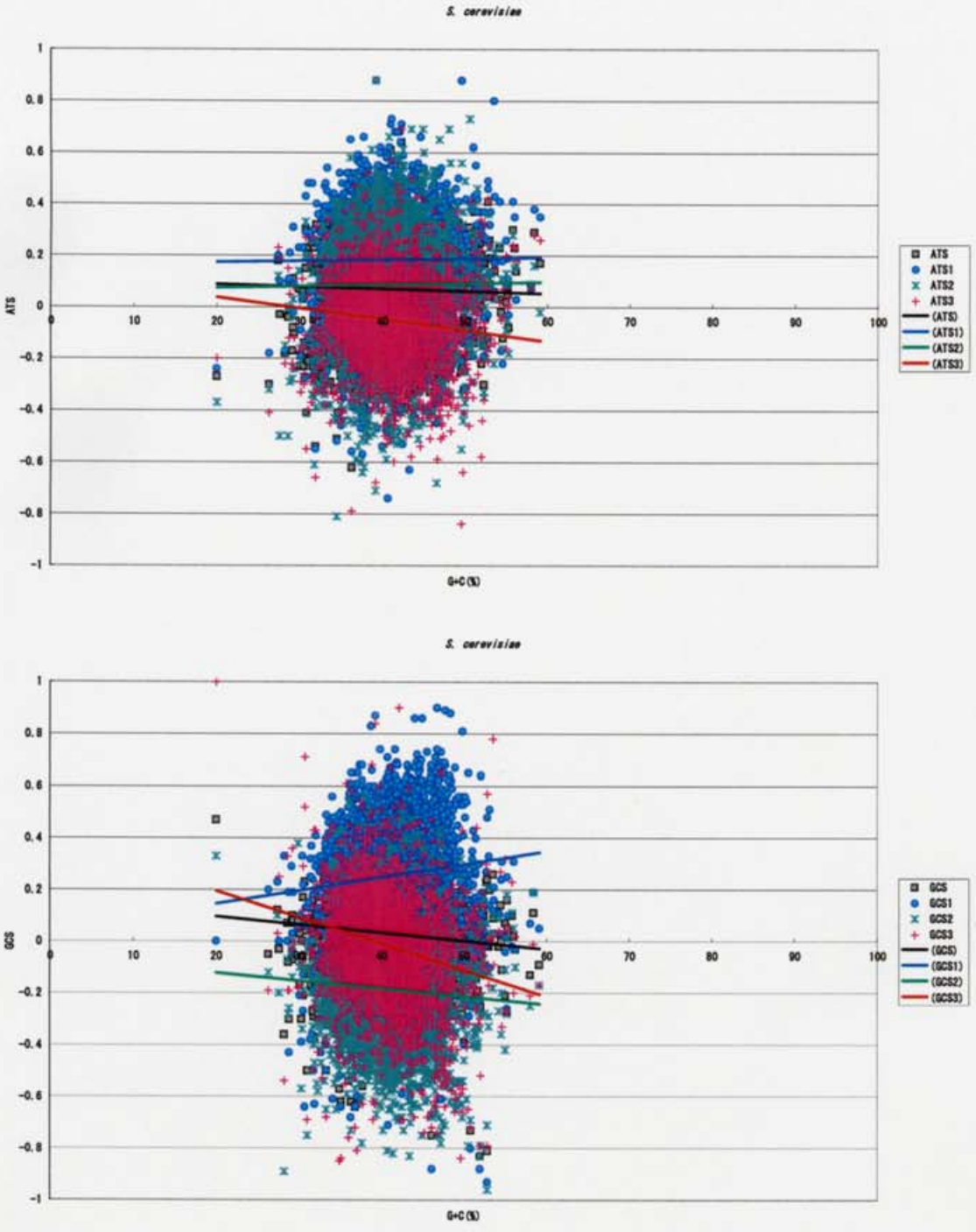


図 3.16 *S.cerevisiae* における CDS の GC 含量と各コドンポジション別 ATS（上図）と GCS（下図）の相関。ATS では、第 1、第 2、第 3 の順であり、GCS では、第 1、第 3、第 2 の順である。

3.3 結論

Skew のパターンからわかることは、複製に関する鎖の性質の入れ替わりである。細菌では、ゲノム全体が 1 つの複製単位になっているために、Skew のパターンから複製開始点が容易に予測できる。真核生物では複製のシステムが複雑であるため、複製単位の予測は難しいが、Skew のパターンの入れ替わりから、リーディング鎖とラギング鎖の入れ替わりが推測可能である。こうしたパターンは種固有的ではなく、同一種内でも染色体によって多様である。いずれにしても、すべてのゲノムにおいて鎖非対称性があることは明らかであり、複製による突然変異は鎖非対称に生じている。

また、CDS における GC 含量との相関では、いずれも ATS では第 1、第 2、第 3 の順であり、GCS では第 1、第 3、第 2 の順である。つまり、第 1 ポジションでは A も G も多く、第 2 では T が、第 3 では G が多い傾向がある。

4. 二連塩基（dinucleotide）のゲノム固有的特性とコドン使用頻度

4.1 背景

ゲノムの DNA 塩基配列はランダムではなく、DNA の高次構造や機能と関係して、高頻度に表れる配列やきわめて頻度の低い配列がある。真核生物のコード領域では、CpG や TpA の二連塩基が低く抑えられており、TpG や CpT が高頻度に出現する（Ohno 1988）。同様に、原核生物やオルガネラ、ウィルスの配列にも、それぞれに種固有的な法則性があるが（Karlin and Burge 1995; Karlin and Marázek 1997）、Karlin and Burge(1995)は、メチル化-デアミ化による C->T 突然変異だけでは、動物のミトコンドリアにおける CpG の抑制は説明できないため、二連塩基に頻度特性はDNAの二重らせんの形態の変形しやすさを決めるスタッキング・エネルギーが関係していると主張している。リン酸結合で連結している２つの塩基に組み合わせによって、スタッキング（π－π）相互作用エネルギーが異なり、DNA の二重らせんの形態の変形しやすさが異なり、TpA は最も柔軟で CpG は最も硬い（Hunter 1993）。そのため、TpA は全体では抑制されているが、TATA-box など DNA の構造を折り曲げる認識配列に用いられており、CpG の抑制はポリ（CG）が左巻きの Z 型 DNA 構造をとることも関係していると考えられている。

コード領域はアミノ酸をコードしているため、コドンによる制約がある。遺伝暗号には縮退があり（表 4.1）、同義コドンの使用頻度は種によって、遺伝子によって異なる。大腸菌や酵母では、翻訳効率と関係して、同義コドンの使用頻度と tRNA 量との間に強い相関がある（Ikemura 1981, 1982）。コドン使用頻度については、アイソコアや GC 含量との関係など、他にも様々な説があるが、一般に CpG や TpA が抑制されているゲノムでは、コード領域においても CpG や UpA を含むコドンの使用頻度が抑制されており、コドン使用頻度は二連塩基の頻度傾向に強く影響を受けている（Nussinov 1981）。

表 4.1 標準遺伝暗号表（縮退別色分け）

コドン	アミノ酸	コドン	アミノ酸	コドン	アミノ酸	コドン	アミノ酸
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

水色　6重縮退；白　4重縮退；緑　3重縮退；黄色　2重縮退；紫　非縮退
赤　開始コドン（非縮退）；青　終止コドン（3重縮退）

4.2　方法

二連塩基は４ｘ４＝１６通りの組み合わせがある。XpY の頻度を f_{XY} とし、塩基 X の頻度を f_X とすると、相対的な頻度傾向は以下のような期待値との比 ρ_{XY} から求められる。

$$\rho_{XY} = \frac{f_{XY}}{f_X f_Y}$$

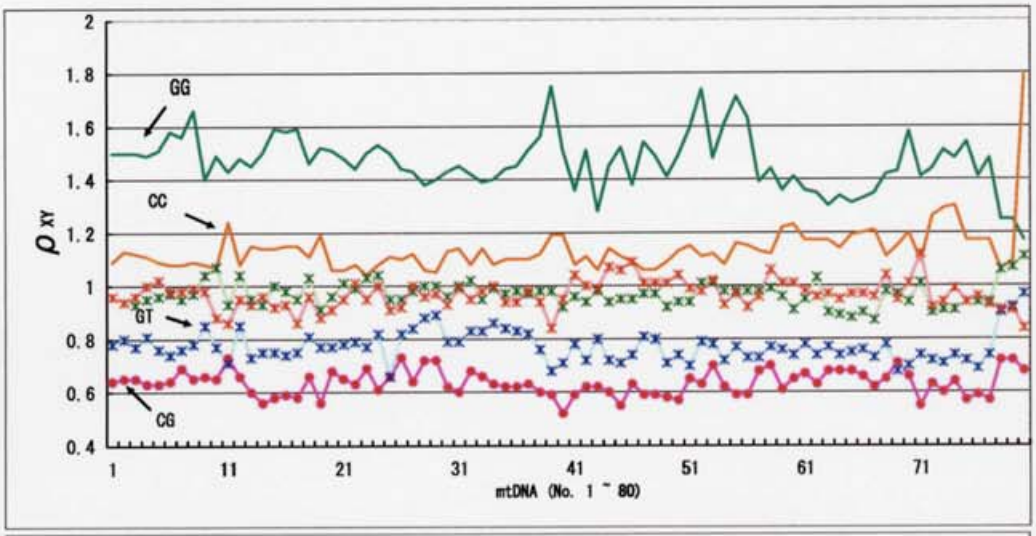
組み合わせの対称性を考慮した１０通りの組み合わせ（ρ^*_{XY}）も考案されているが(Burge et al. 1992)、本研究では鎖の非対称性を考慮して上記の式を用いて、３章の Skew の計算と同様に行った。

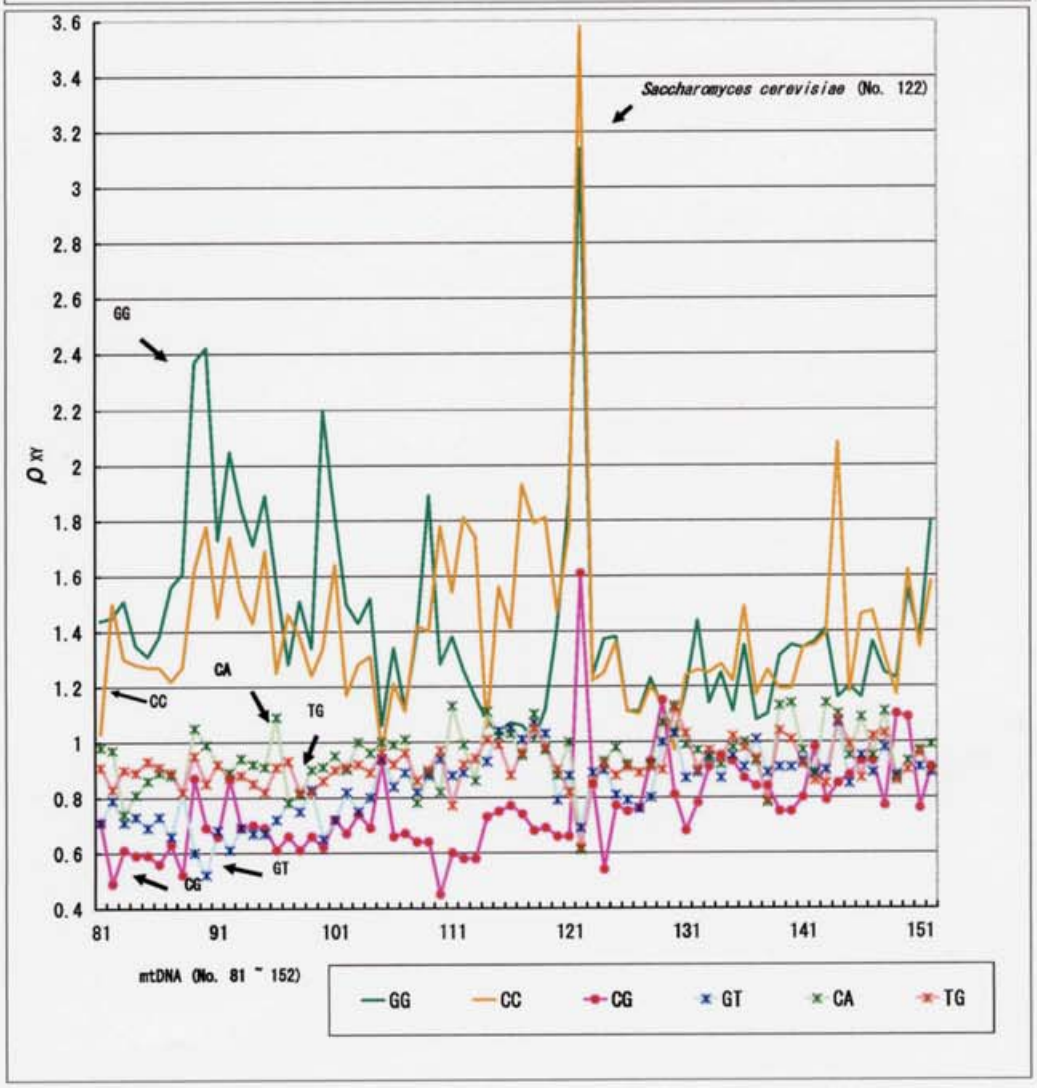
4.3 結果

ミトコンドリア

二連塩基頻度に関して、ミトコンドリア特有の傾向がいくつか観察された（図4.1）。まず、Skewの傾向が配列全体で一貫しており、ATS ≠ 0 またはGCS ≠ 0 が、二連塩基の頻度にも影響し、相補反転の組み合わせ間で頻度に偏りが生じている。例えば、有顎脊椎動物のmtDNAでは、$f_{CA} > f_{TG}$, $f_{AC} > f_{GT}$, $f_{TC} > f_{GA}$, $f_{CT} > f_{AG}$であるが、その他のmtDNAでは逆転している組み合わせもある。つまり、一塩基レベルのSkewは二連塩基にも影響を与えている。次に、ホモ二連塩基GpGやCpCは多くのmtDNAにおいて高頻度であるが、CpGやGpTは低く抑えられている。この傾向は、メチル化-デアミ化による変異とは関係がない。なぜならば、メチル化の機構を持たないショウジョウバエ（*Drosophila*）においても同じ傾向が観察されるからである。また、酵母（*S. cerevisiae*）のmtDNAは異常にGpGとCpCとなる傾向が強い。８６ｋｂのゲノムのGC含量は１７％と低く非常に高ATで、CpGは低く抑えられているが、CとGはクラスターとなって島状に分布している。そして３つ目の特徴は、CpGの減少がTpGとCpAの増加を伴わないことであり、このことはメチル化-デアミ化によるC->T置換では説明できない。Cardon et al. (1994)やKarlin and Burge (1995)は、二連塩基頻度に影響を与える機構がその他に存在することを示唆している。

図 4.1 １５２種のミトコンドリアゲノムにおける二連塩基の頻度/期待値（CpG, CpC, CpG, GpT, CpA, TpG）。脊椎動物（上図）では、いずれも傾向が似ており、その他の種（下図）では多様化している。横軸は、種の分類順（表 A.5）。


GG
CC
GT
CG
ρ XY
mtDNA (No. 1 ~ 80)


Saccharomyces cerevisiae (No. 122)
GG
CC
CA
TG
GT
CG
ρ XY
mtDNA (No. 81 ~ 152)
GG
CC
CG
GT
CA
TG

原核生物

二連塩基の頻度傾向は mtDＮＡとは異なり、相補反転的な組み合わせ同士は同じ傾向を持つ（Burge et al. 1992）。ヘテロ二連塩基だけでなく、ホモ二連塩基についても同様であり、二連塩基の頻度の全体値は鎖対称的である。しかしながら、ウィンドウを作って詳しく調べると、明確な Skew の変化点（複製の開始点または終止点）では、相補反転的な組み合わせが入れ替わっている。二連塩基の偏りも、

$$DNS = \frac{f_{XY} - f_{X'Y'}}{f_{XY} + f_{X'Y'}}$$

として計算できる。ここで、X'Y'は XpY の相補反転である。多くの原核生物において、DNS の値は全ての組み合わせで０に近くなるが、Proteobacteria の γ-subdivision の *Xylella fastidiosa* では、GpT と TpG がその相補反転の ApC と CpA に比べて２０％も多い。ATS と GCS のパターンも、途中で配列が切れたようになっている。

もう一つの驚くべき特徴は、普遍的な性質と思われていた CpG の減少が、いくつかの原核生物では当てはまらないことである。古細菌では、*Halobacterium* sp. NRC-1（GC 含量６８％）が $\rho_{CG} > 1$ であり、真正細菌でも Bacillus グループ（枯草菌など）や Proteobacteria の α, β, γ-subdivision のほとんどの種で $\rho_{CG} > 1$ と CpG が高頻度で観察される。真正細菌では、CpG ではなく TpA や GpT が最も抑制されており、ApC や ApT がそれらに続く。*Ureaplasma urealyticum* は、非常に GC 含量が低く２５．５％であるが、ρ_{TA} は 0.79 であり ρ_{CG} = 0.88 よりわずかに小さい。TpA の抑制はむしろ普遍的であり、例外は古細菌の *Aeropyrum pernix*（ρ_{TA} = 1.21）などわずかである。

いくつかの種では、二連塩基のパターンが局所的に乱れる領域があり、そのような領域では水平転移が生じている可能性が高い。Proteobacteria α-subdivision の *Mesorhizobium loti* では非常に安定した二連塩基のパターンを持つが、５Mb 付近で高頻度な ApT が下がり、抑制されている TpA が増加している。GC 含量もこの付近でちょうど減少しており、何らかの配列変化が生じている。また、Bacillus グループの *Staphylococcus* は、明確な ATS と GCS のパターンを持つが、Skew の変化点で GpC と CpA が急激に上昇し、GpG と CpC が急降下する。これらは、複製開始点に特異的な配列によるものと考えられる。

ヒトゲノム

ヒトゲノムの二連塩基の特徴として、ホモ二連塩基が多いことが挙げられる。そして、パターンは GC 含量の推移を反映しているが、$\rho_{GG} \approx \rho_{CC}$ は $\rho_{AA} \approx \rho_{TT}$ よりも大きく、CpG は期待値のわずか２０～４０％である。ウィンドウサイズを小さくすると、f_{CG} のスパイクが無数に現れるが、これらはメチル化されない CpG アイランドであり、R バンドに集中しているものである（Cross and Bird 1995)。二連塩基の頻度傾向はいずれの染色体でも共通しており、４つのグループに分けられる。まず、ρ＞１であるものは、ApA, TpT, GpG, CpC, TpG, CpA, ApG, CpT である。次に、ρ≈0.8 の ApT, TpA, ApC, GpT と ρ≈0.2 の CpG、そして最後に ρ≈１の GpA, TpC, GpC である(図 3.7 ~ 図 3.9)。CpG の抑制は、mtDNAと同様に TpG, CpA の増加で補填しきれてはいないが、mtDNA と異なり相補反転的な二連塩基の傾向はつねにほぼ同じである（DNS≈0)。

その他の真核生物

それぞれに種固有的な二連塩基のパターンを持つが、真核生物に共通な傾向が多い。近縁種同士は似たパターンを持つものの、必ずしも大きな系統関係を反映しているわけではない。パターンの組み合わせから、ヒトも含めシロイヌナズナ、線虫、酵母の中で、酵母が最も保存的な傾向であるといえる(図 4.2)。

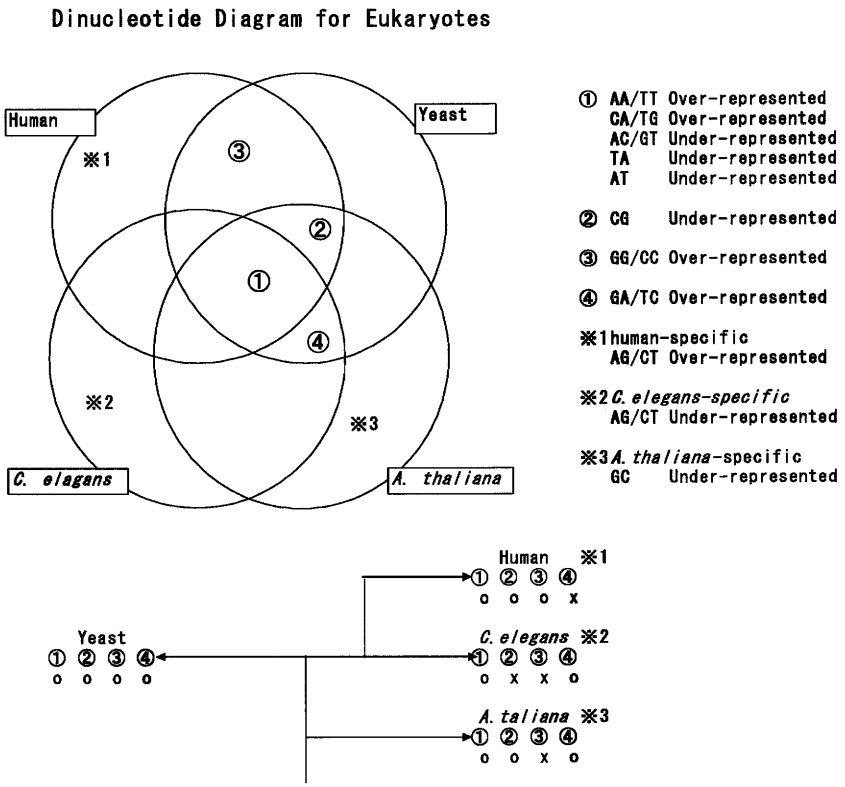


図 4.2 真核生物における二連塩基の頻度傾向。それぞれに種固有的な二連塩基のパターンを持つが、真核生物に共通な傾向が多い。近縁種同士は似たパターンを持つものの、必ずしも大きな系統関係を反映しているわけではない。パターンの組み合わせから、ヒトも含めシロイヌナズナ、線虫、酵母の中で、酵母が最も保存的な傾向であるといえる。

4.4 結論

一塩基レベルの鎖非対称性にかかわらず、それぞれのゲノム配列全体に一貫した種固有的な二連塩基の頻度傾向がある。原核生物から真核生物まであらゆる種においてTpAは少なく、ほぼ普遍的な傾向である。メチル化-デアミ化による突然変異のほかにCpGを抑制する何らかの機構があると考えられ、最も有力な説はDNAの高次構造と関係するスタッキング・エネルギーが原因とする説である。スタッキング・エネルギーとは原子間あるいは分子間の相互作用による結合エネルギーであるため、それぞれの二連塩基の組み合わせで異なる。そして、このエネルギーによる違いがDNAの高次構造の可塑性を決めるが、TpAの組み合わせは最も可塑性が高くCpGは最も硬質であるうえポリ（CG）は左巻き二重らせんのZ型DNA構造をとる。そのため、TpAやCpGが避けられているとされている。そして、その他の二連塩基も含めて１６通りの組み合わせ全てにおいて、ゲノムに一貫した頻度傾向があることも説明できる。また、同義コドンの使用頻度もそれぞれの種で偏りがあり、下等な生物ではｔRNAの存在量とも強い相関があるが、ヒトの遺伝子ではCpGやTpAを含むコドンの使用頻度が抑えられており、ゲノム全体における二連塩基の頻度傾向の影響を強く受けている。その他の真核生物においても同様に、コドンの使用頻度は二連塩基の頻度傾向の影響を強く受けており、二連塩基の組み合わせは基本的に非常に強い性質であるといえる。

5　ゲノム進化に関する考察

生命の誕生や初期進化については様々な仮説があり、いずれも大変興味深いものであるが、残念ながらそれらを実際に確かめることは不可能である。ゲノムも最初はDNAではなくRNAで、熱に対する安定性からDNA配列が一般的になったとされているが、本文でもふれたようにDNAの高次構造は大変可塑的であり、B-型DNAの二重らせんの他にもA-型やZ-型などの分子形態を持つことが、生命活動を行う上で有利であったように思われる。そして、その構造は配列の並びによって決まるため、配列の最も基本的な一塩基レベルや二連塩基レベルで、このように様々なパターンが機能と関係して観察されることは、配列と構造と機能が密接に関連している証拠である。

生命活動における有利性といっても種によって様々であり、そのときそのときの環境や体制そして偶然性によって、普遍的な配列頻度傾向が保存されながらも、現在のような種固有のパターンを持つように至ったのであろう。詳細な解析によって、それらのパターンを生み出す機構が一つ一つ明らかになっていくのかもしれない。そして、それらを全て総合することによって、生命とは何か？あるいは、どのようにして誕生し、どのような進化を遂げていくのか？という問題の理解が深められていくことと思う。

6　結論

本研究の結果から、Skewや二連塩基頻度に関してゲノム配列の数量的特徴づけが明らかとなった。しかしながら、膨大なゲノムの配列データには、そのほかにも様々なパターンが存在しうる。それらのパターンや新たなデータ解析によって、これまでに提唱された分子進化や集団遺伝学における理論の検証や、更なる発展が期待される。

7．引用文献

Beletskii A, Bhagwat AS (1996) Transcription-induced mutations: Increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli.* Proc Natl Acad Sci USA 93:13919-13924

Bernardi G (1995) The human genome: Organization nad evolutionary histoy. Annu Rev Genet 29:445-476

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953-958

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: Tempo and mode of evolution. J Mol Evol 18:225-239

Burge C, Campbell AM, Karlin S (1992) Over- and under-representation of short oligonucleotides in DNA. Proc Natl Acad Sci USA 82:1358-1362

Cardon LR, Burge C, Clayton DA, Karlin S (1994) Pervasive CpG suppression in animal mitochondrial genomes. Proc Natl Acad Sci USA 91:3799-3803

Cross SH, Bird AP (1995) CpG islands and genes. Curr Opin Genet Dev 5:309-314

Felsenstein J (1981) Evolutionary trees fron DNA sequences: a maximum likelihood approach. J Mol Evol 17:368-376

Francino MP, Ochman H (1997) Strand asymmetries in DNA evolution. Trends genet 13:240-245

Freeman JM, Plasterer TN, Smith TF, Mohr SC (1998) Patterns of genome organization in bacteria. Science 279:1827

Fukagawa T, Sugaya K, Matsumoto K, Okumura K, Ando A, Inoko H, Ikemura T (1995) A boundary of long-range G + C% mosaic domains in the human MHC locus: pseudoautosomal boundary-like sequence exists near the boundary. Genomics 25:184-191

Furusawa M, Doi H (1992) Promotion of evolution: Disparity in the frequency of strand-specific misreading between the lagging and leading DNA strands enhances disproportionate accumulation of mutations. J Theor Biol 157:127-133

Furusawa M, Doi H (1998) Asymmetrical DNA replication promotes evolution: disparity theory of evolution. Genetica 102/103:333-347

Gojobori T, Li W-H, Graur D (1982) Pattern of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18:360-369

Graur D (1985) Amino Acid composition and the evolutionary rates of protein-coding genes. J Mol Evol 22:53-62

Grigoriev A (1998) Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res 26:2286-2290

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160-147

Holland J, Spindler K, Horodyski F, Grabau E, Nichol S, vandePol S(1982) Rapid evolution of RNA genomes. Science 215:1577-1585

Hori H and Osawa S (1987) Origin and evolution of organisms as deduced from 5S ribosomal RNA sequences. Mol Biol Evol 4 :445-472

Horiuchi T (1995) Recombinational rescue of the stalled DNA replication fork: a model based on analysis of an Escherichia coli strain with a chromosome region difficult to replicate. J Bacteriol 1995 177(3):783-91

Hughes and Nei (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167-170

Hunter C (1993) Sequence-dependent DNA structure: The role of base stacking interactions. J Mol Biol 230:1025-1054

Ikemura T (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol 15;146(1):1-21

Ikemura T (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol 15;158(4):573-97

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13-34

Ikemura T, Wada K, Aota S (1990) Giant G+C% mosaic structures of the human genome found by arrangement of GenBank human DNA sequences according to genetic positions. Genomics 2:207-216

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian Protein Metabolism III. Academic Press, New York, pp 21-132

Jukes TH, Osawa S (1990) The genetic code in mitochondria and chloroplast. Experientia 46:1117-1126

Jukes TH, Osawa S (1993) Evolutionary changes in the genetic code. Comp Biochem Physiol 106B:489-494

Karlin S, Burge C (1995) Dinucleotide relative abundance extremes: a genome signature. Trends genet 11:283-290

Karlin S, Mrázek J (1997) Compositional differences within and between eukaryotic genomes. Proc Natl Acad Sci USA 94:10227-10232

Kimura M (1968) Genetic variability maintained in a finite population due to neutral and nearly neutral isoalleles. Genet Res Camb 11, 247-69.

Kimura and Ohta(1974) On some principles governing molecular evolution. Proc Natl Acad USA 71:2848-52

Kimura, M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111-120

Kunkel TA (1992) Biological asymmetries and the fidelity of eukaryotic DNA replication. Bioessays 14:303-308

Lecrenier N, Foury F (2000) New features of mitochondrial DNA replication system in yeast and man. Gene 246:37-48

Li W-H, Wu C-I, Luo C-C (1984) Nonrandomeness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21:58-71

Lobry JR (1995) Properties of a general model of DNA evolution under no-strand-bias conditions. J Mol Evol 40:326-330

Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13:660-666

Lopez P, Forterre P, Guyader H, Philippe H (2000) Origin of replication of *Thermotoga maritima*. Trends genet 16:59-60

McLean MJ, Wolfe KH, Devine KM (1998) Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J Mol Evol 47:691-696

Muto A and Osawa S (1987) The guanine and cytocine content of genomic DNA and bacterial evolution. Pro Natl Acad Sci USA 84:166-169

Nei M (1987) *Molecular evolutionary genetics.* Columbia Univ Press

Ohno S (1970) *Evolution by Gene Duplication.* Berlin: Springer-verlag.

Darwin C (1859) *The Origin of Species by Means of Natural Selection.* London:John Murray.

Ohno S (1988) Universal rule for coding sequence construction: TA/CG deficiency–TG/CT excess. Proc Natl Acad Sci USA 85:9630-9634

Osawa S (1994) Evolution of the Genetic Code. Oxford Scientific Publications, Tokyo.

Stryer L(1995) *Biochemistry.* W.H. Freeman and Co.4th ed.

Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40:318-325

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437-460

Takahata N (1985) Population genetics of extranuclear genomes: a model and review. In Ohta T and Aoki K ed. Population Genetics and Molecular Evolution, pp. 195-212

Takahata and Nei (1985) Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110:325-344

Tamura K and Nei M (1993) Estimation of the number of nucleotides substitutions in the control region of mitochondrial DNA in humans and chimpanzee. Mol Biol Evol 10:512-526

Waga S, Stillman B (1994) Anatomy of a DNA replication fork revealed by reconstitution of SV40 DNA replication in vitro. Nature 369:207-212

Whittaker RH (1969) New concepts of kingdoms or organisms. Science 163:150-160

Woese CR and Fox GE (1977) Phylogenetic structure of the prokaryotic domain: Primary kingdoms. Proc Natl Acad Sci USA 74:5088-5090

Wu CI (1991) DNA strand asymmetry. Nature 352:114

Wu C-I and Li W-H (1985) Evidence for higher rates of nucleotides substitution in rodents than in man. Proc Natl Acad Sci USA 82:1741-1745

Wu CI, Maeda N (1987) Inequality in mutation rates of the two strands of DNA. Nature 327:169-170

Zuckerkandle E and Pauling L (1965) Evolutionary divergence and convergenece in proteins. In evolutionary genes and Proteins, Bryoson V and Vogel HJ ed., pp. 97-166. New york: Academic Press.

付録 A. 本文中に関連する表

表 A.1 ミトコンドリア完全ゲノム（３５９）

MITOCHONDRIA

No.	SPECIES	AC NO.	BASES
1	Acanthamoeba castellanii	NC_001637	41591□bp
2	Acropora tenuis	NC_003522	18338□bp
3	Albinaria caerulea	NC_001761	14130□bp
4	Alligator mississippiensis	NC_001922	16646□bp
5	Allocyttus niger	NC_004398	16750□bp
6	Allomyces macrogynus	NC_001715	57473□bp
7	Ancylostoma duodenale	NC_003415	13721□bp
8	Anguilla japonica	NC_002707	16685□bp
9	Anomalopteryx didiformis	NC_002779	16716□bp
10	Anopheles gambiae	NC_002084	15363□bp
11	Anopheles quadrimaculatus A	NC_000875	15455□bp
12	Anoplogaster cornuta	NC_004391	16511□bp
13	Antigonia capros	NC_003191	16508□bp
14	Aphredoderus sayanus	NC_004372	16601□bp
15	Apis mellifera ligustica	NC_001566	16343□bp
16	Apteryx haastii	NC_002782	16816□bp
17	Arabidopsis thaliana	NC_001284	366923□bp
18	Arbacia lixula	NC_001770	15719□bp
19	Arcos sp. KU-149	NC_004413	16534□bp
20	Arctocephalus forsteri	NC_004023	15413□bp
21	Arctoscopus japonicus	NC_002812	16577□bp
22	Arenaria interpres	NC_003712	16725□bp
23	Artemia franciscana	NC_001620	15822□bp
24	Artibeus jamaicensis	NC_002009	16651□bp
25	Ascaris suum	NC_001327	14284□bp
26	Asterina pectinifera	NC_001627	16260□bp
27	Ateleopus japonicus	NC_003178	16650□bp
28	Aulopus japonicus	NC_002674	16653□bp
29	Aythya americana	NC_000877	16616□bp
30	Balaenoptera musculus	NC_001601	16402□bp

31	Balaenoptera physalus	NC_001321	16398□bp
32	Balanoglossus carnosus	NC_001887	15708□bp
33	Bassozetus zenkevitchi	NC_004374	16579□bp
34	Beryx decadactylus	NC_004393	16535□bp
35	Beryx splendens	NC_003188	16529□bp
36	Bombyx mandarina	NC_003395	15928□bp
37	Bombyx mori	NC_002355	15643□bp
38	Bos taurus	NC_001567	16338□bp
39	Branchiostoma floridae	NC_000834	15083□bp
40	Branchiostoma lanceolatum	NC_001912	15076□bp
41	Brugia malayi	NC_004298	13657□bp
42	Buteo buteo	NC_003128	18674□bp
43	Caelorinchus kishinouyei	NC_003169	15942□bp
44	Caenorhabditis elegans	NC_001328	13794□bp
45	Cafeteria roenbergensis	NC_000946	43159□bp
46	Caiman crocodilus	NC_002744	17900□bp
47	Candida albicans	NC_002653	40420□bp
48	Canis familiaris	NC_002008	16727□bp
49	Carangoides armatus	NC_004405	16556□bp
50	Caranx melampygus	NC_004406	16593□bp
51	Carapus bermudensis	NC_004373	16613□bp
52	Carassius auratus	NC_002079	16578□bp
53	Casuarius casuarius	NC_002778	16757□bp
54	Cataetyx rubrirostris	NC_004375	16495□bp
55	Caulophryne pelagica	NC_004383	16657□bp
56	Cavia porcellus	NC_000884	16801□bp
57	Cebus albifrons	NC_002763	16554□bp
58	Cepaea nemoralis	NC_001816	14100□bp
59	Ceratitis capitata	NC_000857	15980□bp
60	Ceratotherium simum	NC_001808	16832□bp
61	Cetostoma regani	NC_004389	16508□bp
62	Chaetosphaeridium globosum	NC_004118	56574□bp
63	Chalinolobus tuberculatus	NC_002626	16818□bp
64	Chauliodus sloani	NC_003159	17814□bp
65	Chaunax abei	NC_004381	16486□bp
66	Chaunax tosaensis	NC_004382	16488□bp

67	Chelonia mydas	NC_000886	16497□bp
68	Chimaera monstrosa	NC_003136	18580□bp
69	Chlamydomonas eugametos	NC_001872	22897□bp
70	Chlamydomonas reinhardtii	NC_001638	15758□bp
71	Chlorophthalmus agassizi	NC_003160	16221□bp
72	Chondrus crispus	NC_001677	25836□bp
73	Chrysemys picta	NC_002073	16866□bp
74	Chrysodidymus synuroideus	NC_002174	34119□bp
75	Chrysomya chloropyga	NC_002697	15837□bp
76	Ciconia boyciana	NC_002196	17622□bp
77	Ciconia ciconia	NC_002197	17347□bp
78	Cochliomyia hominivorax	NC_002660	16022□bp
79	Cololabis saira	NC_003183	16499□bp
80	Conger myriaster	NC_002761	18705□bp
81	Coregonus lavaretus	NC_002646	16737□bp
82	Corvus frugilegus	NC_002069	16932□bp
83	Cottus reinii	NC_004404	16561□bp
84	Coturnix japonica	NC_003408	16697□bp
85	Crassostrea gigas	NC_001276	18224□bp
86	Crenimugil crenilabis	NC_003170	16019□bp
87	Crioceris duodecimpunctata	NC_003372	15880□bp
88	Crossostoma lacustre	NC_001727	16558□bp
89	Cryptococcus neoformans var. grubii	NC_004336	24874□bp
90	Cyanidioschyzon merolae	NC_000887	32211□bp
91	Cynocephalus variegatus	NC_004031	16748□bp
92	Cyprinus carpio	NC_001606	16575□bp
93	Dactyloptena peterseni	NC_003194	16717□bp
94	Dactyloptena tiltoni	NC_004402	16751□bp
95	Danacetichthys galathenus	NC_003185	16555□bp
96	Danio rerio	NC_002333	16596□bp
97	Daphnia pulex	NC_000844	15333□bp
98	Dasypus novemcinctus	NC_001821	17056□bp
99	Diaphus splendidus	NC_003164	15985□bp
100	Dictyostelium discoideum	NC_000895	55564□bp
101	Didelphis virginiana	NC_001610	17084□bp
102	Dinodon semicarinatus	NC_001945	17191□bp

103	Dinornis giganteus	NC_002672	17070□bp
104	Diplacanthopoma brachysoma	NC_004376	16495□bp
105	Diplophos taenia	NC_002647	16418□bp
106	Dogania subplana	NC_002780	17289□bp
107	Dromaius novaehollandiae	NC_002784	16711□bp
108	Drosophila melanogaster	NC_001709	19517□bp
109	Drosophila yakuba	NC_001322	16019□bp
110	Dugong dugon	NC_003314	16850□bp
111	Echinococcus multilocularis	NC_000928	13738□bp
112	Echinops telfairi	NC_002631	16555□bp
113	Echinosorex gymnura	NC_002808	17088□bp
114	Elassoma evergladei	NC_003175	15780□bp
115	Eleotris acanthopoma	NC_004415	16522□bp
116	Emeus crassus	NC_002673	17061□bp
117	Emmelichthys struhsakeri	NC_004407	16502□bp
118	Enedrias crassispina	NC_004410	16522□bp
119	Engraulis japonicus	NC_003097	16675□bp
120	Eptatretus burgeri	NC_002807	17168□bp
121	Equus asinus	NC_001788	16670□bp
122	Equus caballus	NC_001640	16660□bp
123	Erinaceus europaeus	NC_002080	17447□bp
124	Eudromia elegans	NC_002772	15302□bp
125	Eumeces egregius	NC_000888	17407□bp
126	Eumetopias jubatus	NC_004030	16639□bp
127	Eutaeniophorus sp. 033-Miya	NC_004390	16508□bp
128	Exocoetus volitans	NC_003184	16527□bp
129	Falco peregrinus	NC_000878	18068□bp
130	Fasciola hepatica	NC_002546	14462□bp
131	Felis catus	NC_001700	17009□bp
132	Florometra serratissima	NC_001878	16005□bp
133	Gadus morhua	NC_002081	16696□bp
134	Gallus gallus	NC_001323	16775□bp
135	Gambusia affinis	NC_004388	16614□bp
136	Gasterosteus aculeatus	NC_003174	15742□bp
137	Gonostoma gracile	NC_002574	16436□bp
138	Gorilla gorilla	NC_001645	16364□bp

139	Gymnothorax kidako	NC_004417	16579□bp
140	Haematopus ater	NC_003713	16589□bp
141	Halichoerus grypus	NC_001602	16797□bp
142	Halocynthia roretzi	NC_002177	14771□bp
143	Harpadon microchir	NC_003161	16061□bp
144	Helicolenus hilgendorfi	NC_003195	16728□bp
145	Heterodontus francisci	NC_003137	16708□bp
146	Heterodoxus macropus	NC_002651	14670□bp
147	Hippopotamus amphibius	NC_000889	16407□bp
148	Homo sapiens	NC_001807	16571□bp
149	Hoplostethus japonicus	NC_003187	16528□bp
150	Hyaloraphidium curvatum	NC_003048	29593□bp
151	Hylobates lar	NC_002082	16472□bp
152	Hymenolepis diminuta	NC_002767	13900□bp
153	Hypoatherina tsurugae	NC_004386	16566□bp
154	Hypocrea jecorina	NC_003388	42130□bp
155	Hypoptychus dybowskii	NC_004400	16479□bp
156	Ictalurus punctatus	NC_003489	16497□bp
157	Iguana iguana	NC_002793	16633□bp
158	Ijimaia dofleini	NC_003179	16645□bp
159	Indostomus paradoxus	NC_004401	16152□bp
160	Isoodon macrourus	NC_002746	16852□bp
161	Ixodes hexagonus	NC_002010	14539□bp
162	Ixodes persulcatus	NC_004370	14539□bp
163	Katharina tunicata	NC_001636	15532□bp
164	Lama pacos	NC_002504	16652□bp
165	Laminaria digitata	NC_004024	38007□bp
166	Lampetra fluviatilis	NC_001131	16159□bp
167	Lampris guttatus	NC_003165	15598□bp
168	Laqueus rubellus	NC_002322	14017□bp
169	Latimeria chalumnae	NC_001804	16407□bp
170	Lemur catta	NC_004025	17036□bp
171	Lepidosiren paradoxa	NC_003342	16403□bp
172	Lepus europaeus	NC_004028	17734□bp
173	Limulus polyphemus	NC_003057	14985□bp
174	Lithobius forficatus	NC_002629	15695□bp

175	Locusta migratoria	NC_001712	15722□bp
176	Loligo bleekeri	NC_002507	17211□bp
177	Lophius americanus	NC_004380	16479□bp
178	Lota lota	NC_004379	16527□bp
179	Loxodonta africana	NC_000934	16866□bp
180	Lumbricus terrestris	NC_001673	14998□bp
181	Lycodes toyamensis	NC_004409	16697□bp
182	Macaca sylvanus	NC_002764	16586□bp
183	Macropus robustus	NC_001794	16896□bp
184	Macroscelides proboscideus	NC_004026	16641□bp
185	Malawimonas jakobiformis	NC_002553	47328□bp
186	Manis tetradactyla	NC_004027	16571□bp
187	Marchantia polymorpha	NC_001660	186609□bp
188	Mastacembelus favus	NC_003193	16498□bp
189	Melanocetus murrayi	NC_004384	16758□bp
190	Melanonus zugmayeri	NC_004378	16564□bp
191	Melanotaenia lacustris	NC_004385	16487□bp
192	Mertensiella luschani	NC_002756	16650□bp
193	Metridium senile	NC_000933	17443□bp
194	Monocentris japonicus	NC_004392	16567□bp
195	Monopterus albus	NC_003192	16622□bp
196	Monosiga brevicollis	NC_004309	76568□bp
197	Mugil cephalus	NC_003182	16685□bp
198	Muntiacus reevesi	NC_004069	16354□bp
199	Mus musculus	NC_001569	16295□bp
200	Mustelus manazo	NC_000890	16707□bp
201	Myctophum affine	NC_003163	16239□bp
202	Myoxus glis	NC_001892	16602□bp
203	Myripristis berndti	NC_003189	16531□bp
204	Myxine glutinosa	NC_002639	18909□bp
205	Naegleria gruberi	NC_002573	49843□bp
206	Narceus annularus	NC_003343	14868□bp
207	Necator americanus	NC_003416	13604□bp
208	Neoceratodus forsteri	NC_003127	16572□bp
209	Neocyttus rhomboidalis	NC_004399	16745□bp
210	Neoscopelus microchir	NC_003180	16686□bp

211	Nycticebus coucang	NC_002765	16764□bp
212	Ochotona collaris	NC_003033	16968□bp
213	Ochromonas danica	NC_002571	41035□bp
214	Odobenus rosmarus rosmarus	NC_004029	16565□bp
215	Onchocerca volvulus	NC_001861	13747□bp
216	Oncorhynchus mykiss	NC_001717	16642□bp
217	Oncorhynchus tshawytscha	NC_002980	16644□bp
218	Ornithorhynchus anatinus	NC_000891	17019□bp
219	Orycteropus afer	NC_002078	16816□bp
220	Oryctolagus cuniculus	NC_001913	17245□bp
221	Oryzias latipes	NC_004387	16714□bp
222	Osteoglossum bicirrhosum	NC_003095	16006□bp
223	Ostichthys japonicus	NC_004394	16541□bp
224	Ostrinia furnacalis	NC_003368	14536□bp
225	Ostrinia nubilalis	NC_003367	14535□bp
226	Ovis aries	NC_001941	16616□bp
227	Pagrus major	NC_003196	17031□bp
228	Pagurus longicarpus	NC_003058	15630□bp
229	Pan paniscus	NC_001644	16563□bp
230	Pan troglodytes	NC_001643	16554□bp
231	Pantodon buchholzi	NC_003096	15845□bp
232	Panulirus japonicus	NC_004251	15717□bp
233	Papio hamadryas	NC_001992	16521□bp
234	Paracentrotus lividus	NC_001572	15696□bp
235	Paragonimus westermani	NC_002354	14965□bp
236	Paralichthys olivaceus	NC_002386	17090□bp
237	Paramecium aurelia	NC_001324	40469□bp
238	Parazen pacificus	NC_004396	16848□bp
239	Pedinomonas minor	NC_000892	25137□bp
240	Pelomedusa subrufa	NC_001947	16787□bp
241	Penaeus monodon	NC_002184	15984□bp
242	Percopsis transmontana	NC_003168	16079□bp
243	Petromyzon marinus	NC_001626	16201□bp
244	Petroscirtes breviceps	NC_004411	16680□bp
245	Phoca vitulina	NC_001325	16826□bp
246	Physarum polycephalum	NC_002508	62862□bp

247	Physeter catodon	NC_002503	16428□bp
248	Physiculus japonicus	NC_004377	16999□bp
249	Phytophthora infestans	NC_002387	37957□bp
250	Pichia canadensis	NC_001762	27694□bp
251	Plasmodium falciparum	NC_002375	5967□bp
252	Plasmodium reichenowi	NC_002235	5966□bp
253	Platichthys bicoloratus	NC_003176	15973□bp
254	Platynereis dumerilii	NC_000931	15619□bp
255	Plecoglossus altivelis	NC_002734	16537□bp
256	Podospora anserina	NC_001329	94192□bp
257	Polymixia japonica	NC_002648	16481□bp
258	Polymixia lowei	NC_003181	16473□bp
259	Polyodon spathula	NC_004419	16512□bp
260	Polypterus ornatipinnis	NC_001778	16624□bp
261	Polypterus senegalus senegalus	NC_004418	16627□bp
262	Pongo pygmaeus	NC_001646	16389□bp
263	Pongo pygmaeus abelii	NC_002083	16499□bp
264	Poromitra oscitans	NC_003172	16387□bp
265	Porphyra purpurea	NC_002007	36753□bp
266	Protopterus dolloi	NC_001708	16646□bp
267	Prototheca wickerhamii	NC_001613	55328□bp
268	Pterocaesio tile	NC_004408	16496□bp
269	Pterocnemia pennata	NC_002783	16747□bp
270	Pteropus dasymallus	NC_002612	16705□bp
271	Pteropus scapulatus	NC_002619	16741□bp
272	Pupa strigosa	NC_002176	14189□bp
273	Pylaiella littoralis	NC_003055	58507□bp
274	Pyrocoelia rufa	NC_003970	17739□bp
275	Raja radiata	NC_000893	16783□bp
276	Rana nigromaculata	NC_002805	17804□bp
277	Ranodon sibiricus	NC_004021	16418□bp
278	Rattus norvegicus	NC_001665	16300□bp
279	Reclinomonas americana	NC_001823	69034□bp
280	Rhea americana	NC_000846	16714□bp
281	Rhinoceros unicornis	NC_001779	16829□bp
282	Rhipicephalus sanguineus	NC_002074	14710□bp

283	Rhizophydium sp. 136	NC_003053	68834□bp
284	Rhodomonas salina	NC_002572	48063□bp
285	Rhyacichthys aspro	NC_004414	16518□bp
286	Rivulus marmoratus	NC_003290	17329□bp
287	Roboastra europaea	NC_004321	14472□bp
288	Rondeletia loricata	NC_003186	16530□bp
289	Saccharomyces castellii	NC_003920	25753□bp
290	Saccharomyces cerevisiae	NC_001224	85779□bp
291	Salarias fasciatus	NC_004412	16496□bp
292	Salmo salar	NC_001960	16665□bp
293	Salvelinus alpinus	NC_000861	16659□bp
294	Salvelinus fontinalis	NC_000860	16624□bp
295	Sardinops melanostictus	NC_002616	16881□bp
296	Sargocentron rubrum	NC_004395	16526□bp
297	Satyrichthys amiscus	NC_004403	16526□bp
298	Saurida undosquamis	NC_003162	15737□bp
299	Scaphirhynchus cf. albus	NC_004420	16493□bp
300	Scenedesmus obliquus	NC_002254	42781□bp
301	Schistosoma japonicum	NC_002544	14085□bp
302	Schistosoma mansoni	NC_002545	14415□bp
303	Schistosoma mekongi	NC_002529	14072□bp
304	Schizophyllum commune	NC_003049	49704□bp
305	Schizosaccharomyces japonicus	NC_004332	80059□bp
306	Schizosaccharomyces octosporus	NC_004312	44227□bp
307	Schizosaccharomyces pombe	NC_001326	19431□bp
308	Sciurus vulgaris	NC_002369	16507□bp
309	Scopelogadus mizolepis	NC_003171	16375□bp
310	Scyliorhinus canicula	NC_001950	16697□bp
311	Smithornis sharpei	NC_000879	17344□bp
312	Soriculus fumidus	NC_003040	17488□bp
313	Spizellomyces punctatus	NC_003052	58830□bp
314	Spizellomyces punctatus	NC_003060	1136□bp
315	Spizellomyces punctatus	NC_003061	1381□bp
316	Squalus acanthias	NC_002012	16738□bp
317	Stephanolepis cirrhifer	NC_003177	16306□bp
318	Strongylocentrotus purpuratus	NC_001453	15650□bp

319	Struthio camelus	NC_002785	16595□bp
320	Sufflamen fraenatus	NC_004416	16584□bp
321	Sus scrofa	NC_000845	16613□bp
322	Tachyglossus aculeatus	NC_003321	16360□bp
323	Taenia crassiceps	NC_002547	13503□bp
324	Taenia solium	NC_004022	13709□bp
325	Takifugu rubripes	NC_004299	16447□bp
326	Talpa europaea	NC_002391	16884□bp
327	Tamandua tetradactyla	NC_004032	16395□bp
328	Tarsius bancanus	NC_002811	16927□bp
329	Terebratalia transversa	NC_003086	14291□bp
330	Terebratulina retusa	NC_000941	15451□bp
331	Tetrahymena pyriformis	NC_000862	47296□bp
332	Tetrahymena thermophila	NC_003029	47577□bp
333	Tetrodontophora bielanensis	NC_002735	15455□bp
334	Thrips imaginis	NC_004371	15407□bp
335	Thryonomys swinderianus	NC_002658	16626□bp
336	Thyropygus sp. DVL-2001	NC_003344	15133□bp
337	Tigriopus japonicus	NC_003979	14628□bp
338	Tinamus major	NC_002781	16702□bp
339	Trachipterus trachypterus	NC_003166	16162□bp
340	Trachurus japonicus	NC_002813	16559□bp
341	Triatoma dimidiata	NC_002609	17019□bp
342	Tribolium castaneum	NC_003081	15881□bp
343	Trichinella spiralis	NC_002681	16706□bp
344	Trichosurus vulpecula	NC_003039	17191□bp
345	Tupaia belangeri	NC_002521	16754□bp
346	Typhlonectes natans	NC_002471	17005□bp
347	Ursus americanus	NC_003426	16841□bp
348	Ursus arctos	NC_003427	17020□bp
349	Ursus maritimus	NC_003428	17017□bp
350	Venerupis (Ruditapes) philippinarum	NC_003354	22676□bp
351	Vidua chalybeata	NC_000880	16895□bp
352	Volemys kikuchii	NC_003041	16312□bp
353	Vombatus ursinus	NC_003322	16996□bp
354	Xenopus laevis	NC_001573	17553□bp

355	Yarrowia lipolytica	NC_002659	47916□bp
356	Zenion japonicum	NC_004397	16843□bp
357	Zenopsis nebulosus	NC_003173	16065□bp
358	Zeus faber	NC_003190	16715□bp
359	Zu cristatus	NC_003167	15987□bp

表 A.2 古細菌完全ゲノム（１５）

ARCHAEA

NO.	SPECIES	AC NO.	BASES
1	Aeropyrum pernix	NC_000854	1669695□bp
2	Archaeoglobus fulgidus	NC_000917	2178400□bp
3	Halobacterium sp. NRC-1	NC_002607	2014239□bp
4	Methanococcus jannaschii	NC_000909	1664970□bp
5	Methanopyrus kandleri AV19	NC_003551	1694969□bp
6	Methanosarcina acetivorans str. C2A	NC_003552	5751492□bp
7	Methanosarcina mazei Goe1	NC_003901	4096345□bp
8	Methanothermobacter thermautotrophicus str. Delta H	NC_000916	1751377□bp
9	Pyrobaculum aerophilum	NC_003364	2222430□bp
10	Pyrococcus abyssi	NC_000868	1765118□bp
11	Pyrococcus furiosus DSM 3638	NC_003413	1908256□bp
12	Pyrococcus horikoshii	NC_000961	1738505□bp
13	Sulfolobus solfataricus	NC_002754	2992245□bp
14	Thermoplasma acidophilum	NC_002578	1564906□bp
15	Thermoplasma volcanium	NC_002689	1584804□bp

表 A.3 細菌完全ゲノム（９０）

BACTERIA

NO.	SPECIES	AC NO.	BASES
1	Agrobacterium tumefaciens str. C58 (Cereon)	NC_003062	2841581□bp
2	Agrobacterium tumefaciens str. C58 (Cereon)	NC_003063	2074782□bp
3	Agrobacterium tumefaciens str. C58 (U. Washington)	NC_003304	2841490□bp
4	Agrobacterium tumefaciens str. C58 (U. Washington)	NC_003305	2075560□bp
5	Aquifex aeolicus	NC_000918	1551335□bp

6	Bacillus anthracis str. A2012	NC_003995	5093554□bp
7	Bacillus halodurans	NC_002570	4202353□bp
8	Bacillus subtilis	NC_000964	4214814□bp
9	Bifidobacterium longum NCC2705	NC_004307	2256646□bp
10	Borrelia burgdorferi	NC_001318	910724□bp
11	Brucella melitensis	NC_003317	2117144□bp
12	Brucella melitensis	NC_003318	1177787□bp
13	Brucella suis 1330	NC_004310	2107792□bp
14	Brucella suis 1330	NC_004311	1207381□bp
15	Buchnera aphidicola str. Sg (Schizaphis graminum)	NC_004061	641454□bp
16	Buchnera sp. APS	NC_002528	640681□bp
17	Campylobacter jejuni	NC_002163	1641481□bp
18	Caulobacter crescentus CB15	NC_002696	4016947□bp
19	Chlamydia muridarum	NC_002620	1069411□bp
20	Chlamydia trachomatis	NC_000117	1042519□bp
21	Chlamydophila pneumoniae AR39	NC_002179	1229858□bp
22	Chlamydophila pneumoniae CWL029	NC_000922	1230230□bp
23	Chlamydophila pneumoniae J138	NC_002491	1226565□bp
24	Chlorobium tepidum TLS	NC_002932	2154946□bp
25	Clostridium acetobutylicum	NC_003030	3940880□bp
26	Clostridium perfringens	NC_003366	3031430□bp
27	Corynebacterium glutamicum ATCC 13032	NC_003450	3309401□bp
28	Deinococcus radiodurans	NC_001263	2648638□bp
29	Deinococcus radiodurans	NC_001264	412348□bp
30	Escherichia coli CFT073	NC_004431	5231428□bp
31	Escherichia coli K12	NC_000913	4639221□bp
32	Escherichia coli O157:H7	NC_002695	5498450□bp
33	Escherichia coli O157:H7 EDL933	NC_002655	5528445□bp
34	Fusobacterium nucleatum subsp. nucleatum ATCC 25586	NC_003454	2174500□bp
35	Haemophilus influenzae Rd	NC_000907	1830138□bp
36	Helicobacter pylori 26695	NC_000915	1667867□bp
37	Helicobacter pylori J99	NC_000921	1643831□bp
38	Lactococcus lactis subsp. lactis	NC_002662	2365589□bp
39	Leptospira interrogans serovar lai str. 56601	NC_004342	4332241□bp

40	Leptospira interrogans serovar lai str. 56601	NC_004343	358943□bp
41	Listeria innocua	NC_003212	3011208□bp
42	Listeria monocytogenes EGD-e	NC_003210	2944528□bp
43	Mesorhizobium loti	NC_002678	7036074□bp
44	Mycobacterium leprae	NC_002677	3268203□bp
45	Mycobacterium tuberculosis CDC1551	NC_002755	4403836□bp
46	Mycobacterium tuberculosis H37Rv	NC_000962	4411529□bp
47	Mycoplasma genitalium	NC_000908	580074□bp
48	Mycoplasma pneumoniae	NC_000912	816394□bp
49	Mycoplasma pulmonis	NC_002771	963879□bp
50	Neisseria meningitidis MC58	NC_003112	2272351□bp
51	Neisseria meningitidis Z2491	NC_003116	2184406□bp
52	Nostoc sp. PCC 7120	NC_003272	6413771□bp
53	Oceanobacillus iheyensis	NC_004193	3630528□bp
54	Pasteurella multocida	NC_002663	2257487□bp
55	Pseudomonas aeruginosa	NC_002516	6264403□bp
56	Ralstonia solanacearum	NC_003295	3716413□bp
57	Ralstonia solanacearum	NC_003296	2094509□bp
58	Rickettsia conorii	NC_003103	1268755□bp
59	Rickettsia prowazekii	NC_000963	1111523□bp
60	Salmonella enterica subsp. enterica serovar Typhi	NC_003198	4809037□bp
61	Salmonella typhi	NC_002305	180461□bp
62	Salmonella typhimurium LT2	NC_003197	4857432□bp
63	Shewanella oneidensis MR-1	NC_004347	4969803□bp
64	Shewanella oneidensis MR-1	NC_004349	161613□bp
65	Shigella flexneri 2a str. 301	NC_004337	4607203□bp
66	Sinorhizobium meliloti	NC_003047	3654135□bp
67	Staphylococcus aureus subsp. aureus MW2	NC_003923	2820462□bp
68	Staphylococcus aureus subsp. aureus Mu50	NC_002758	2878040□bp
69	Staphylococcus aureus subsp. aureus N315	NC_002745	2813641□bp
70	Streptococcus agalactiae 2603V/R	NC_004116	2160267□bp
71	Streptococcus mutans UA159	NC_004350	2030921□bp
72	Streptococcus pneumoniae R6	NC_003098	2038615□bp
73	Streptococcus pneumoniae TIGR4	NC_003028	2160837□bp
74	Streptococcus pyogenes M1 GAS	NC_002737	1852441□bp

75	Streptococcus pyogenes MGAS315	NC_004070	1900521□bp
76	Streptococcus pyogenes MGAS8232	NC_003485	1895017□bp
77	Synechocystis sp. PCC 6803	NC_000911	3573470□bp
78	Thermoanaerobacter tengcongensis	NC_003869	2689445□bp
79	Thermosynechococcus elongatus BP-1	NC_004113	2593857□bp
80	Thermotoga maritima	NC_000853	1860725□bp
81	Treponema pallidum	NC_000919	1138011□bp
82	Ureaplasma urealyticum	NC_002162	751719□bp
83	Vibrio cholerae	NC_002505	2961149□bp
84	Vibrio cholerae	NC_002506	1072315□bp
85	Wigglesworthia brevipalpis	NC_004344	697721□bp
86	Xanthomonas axonopodis pv. citri str. 306	NC_003919	5175554□bp
87	Xanthomonas campestris pv. campestris str. ATCC 33913	NC_003902	5076188□bp
88	Xylella fastidiosa 9a5c	NC_002488	2679306□bp
89	Yersinia pestis	NC_003143	4653728□bp
90	Yersinia pestis KIM	NC_004088	4600755□bp

表 A.4 ９種の真核生物(１～３はドラフト、４～９は完全ゲノム)
EUKARYOTES

NO.	SPECIES	CHR	LINEAGE
1	*Homo sapiens*	1 ~22 X Y	Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates
2	*Mus musclus*	1~19 X Y	Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Rodentia
3	*Drosophila melanogaster*	X	Metazoa; Arthropoda; Hexapoda; Insecta
4	*Caenorhabditis elegans*	1~5 X	Metazoa; Nematoda
5	*Arabidopsis thaliana*	1~5	Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta
6	*Plasmodium falciparum*	1~14	Eukaryota; Alveolata; Apicomplexa; Haemosporida; Plasmodium
7	*Saccharomyces cerevisiae*	1~16	Fungi; Ascomycota; Saccharomycotina

8	*Schizosaccharomyces pombe*	1~3	Fungi; Ascomycota; Schizosaccharomycetes
9	*Encephalitozoon cuniculi*	1~11	Fungi; Microsporidia; Unikaryonidae; Encephalitozoon

表 A.5 152 種のミトコンドリア完全ゲノムにおける ATS と GCS

	species	bp	G+C [a]	G [b]	ATS [c]	GCS [c]
1	Homo sapiens	16569	44.4	2	0.11	- 0.41
2	Pan paniscus	16563	43.4	2	0.11	- 0.41
3	Pan troglodytes	16554	43.7	2	0.11	- 0.41
4	Gorilla gorilla	16364	43.9	2	0.10	- 0.40
5	Pongo pygmaeus	16389	45.7	2	0.12	- 0.42
6	Pongo pygmaeus abelii	16499	45.9	2	0.13	- 0.43
7	Hylobates lar	16472	45.5	2	0.12	- 0.40
8	Papio hamadryas	16521	43.7	2	0.12	- 0.40
9	Tupaia belangeri	16754	40.8	2	0.10	- 0.29
10	Dasypus novemcinctus	17056	38.9	2	0.13	- 0.34
11	Erinaceus europaeus	17447	32.6	2	0.01	- 0.23
12	Talpa europaea	16884	38.9	2	0.12	- 0.26
13	Oryctolagus cuniculus	17245	40.2	2	0.05	- 0.32
14	Sciurus vulgaris	16507	37.0	2	0.02	- 0.32
15	Mus musculus	16295	36.7	2	0.09	- 0.33
16	Rattus norvegicus	16300	38.7	2	0.11	- 0.36
17	Myoxus glis	16602	36.2	2	0.02	- 0.30
18	Cavia porcellus	16801	39.3	2	0.06	- 0.26
19	Artibeus jamaicensis	16651	37.9	2	0.04	- 0.31
20	Loxodonta africana	16866	38.8	2	0.07	- 0.30
21	Equus asinus	16670	42.1	2	0.12	- 0.37
22	Equus caballus	16660	42.0	2	0.11	- 0.36
23	Ceratotherium simum	16832	40.9	2	0.13	- 0.37
24	Rhinoceros unicornis	16829	40.2	2	0.12	- 0.37
25	Orycteropus afer	16816	38.1	2	0.07	- 0.34
26	Canis familiaris	16728	39.7	2	0.05	- 0.29
27	Felis catus	17009	40.3	2	0.09	- 0.30
28	Halichoerus grypus	16797	41.7	2	0.13	- 0.32
29	Phoca vitulina	16826	41.7	2	0.13	- 0.32
30	Bos taurus	16338	39.4	2	0.10	- 0.32
31	Ovis aries	16616	38.9	2	0.10	- 0.33
32	Sus scrofa	16613	39.5	2	0.15	- 0.33
33	Lama pacos	16652	40.9	2	0.08	- 0.29

34	Hippopotamus amphibius	16407	42.6	2	0.14	- 0.34
35	Balaenoptera musculus	16402	40.7	2	0.10	- 0.36
36	Balaenoptera physalus	16398	40.6	2	0.10	- 0.34
37	Physeter catodon	16428	43.1	2	0.12	- 0.38
38	Macropus robustus	16896	39.2	2	0.09	- 0.34
39	Didelphis virginiana	17084	33.2	2	0.06	- 0.27
40	Ornithorhynchus anatinus	17019	37.1	2	0	- 0.27
41	Corvus frugilegus	16932	44.3	2	0.10	- 0.34
42	Smithornis sharpei	17344	45.2	2	0.10	- 0.44
43	Vidua chalybeata	16895	45.8	2	0.15	- 0.35
44	Falco peregrinus	18068	44.4	2	0.18	- 0.39
45	Gallus gallus	16775	46.0	2	0.12	- 0.41
46	Aythya americana	16616	48.4	2	0.14	- 0.35
47	Ciconia boyciana	17622	46.3	2	0.15	- 0.38
48	Ciconia ciconia	17347	46.3	2	0.14	- 0.38
49	Rhea americana	16714	46.9	2	0.08	- 0.37
50	Struthio camelus	16591	44.7	2	0.10	- 0.36
51	Alligator mississippiensis	16646	43.0	2	0.10	- 0.37
52	Dinodon semicarinatus	17191	39.9	2	0.16	- 0.39
53	Eumeces egregius	17407	44.2	2	0.09	- 0.3
54	Chelonia mydas	16497	39.5	2	0.17	- 0.39
55	Chrysemys picta	16866	38.8	2	0.13	- 0.34
56	Pelomedusa subrufa	16787	38.7	2	0.11	- 0.37
57	Xenopus laevis	17553	37.0	2	0.05	- 0.27
58	Typhlonectes natans	17005	45.1	2	0.10	- 0.29
59	Carassius auratus	16578	42.6	2	0.09	- 0.24
60	Crossostoma lacustre	16558	45.5	2	0.08	- 0.26
61	Cyprinus carpio	16575	43.3	2	0.12	- 0.27
62	Danio rerio	16890	39.8	2	0.06	- 0.20
63	Oncorhynchus mykiss	16642	46.0	2	0.03	- 0.26
64	Salmo salar	16665	45.2	2	0.04	- 0.28
65	Salvelinus alpinus	16659	45.5	2	0.03	- 0.25
66	Salvelinus fontinalis	16624	45.2	2	0.03	- 0.25
67	Gadus morhua	16696	42.4	2	- 0.03	- 0.21
68	Paralichthys olivaceus	17090	46.5	2	0.02	- 0.28
69	Polypterus ornatipinnis	16624	39.8	2	0.07	- 0.29

70	Protopterus dolloi	16646	42.2	2	0	- 0.25
71	Latimeria chalumnae	16407	41.7	2	0.18	- 0.28
72	Mustelus manazo	16707	38.3	2	0	- 0.27
73	Scyliorhinus canicula	16697	38.0	2	- 0.01	- 0.26
74	Squalus acanthias	16738	38.8	2	0.01	- 0.26
75	Raja radiata	16783	40.3	2	0.02	- 0.29
76	Lampetra fluviatilis	16159	38.6	2	0.03	- 0.26
77	Petromyzon marinus	16201	37.3	2	0.03	- 0.28
78	Branchiostoma floridae	15083	37.3	5	- 0.14	0.15
79	Branchiostoma lanceolatum	15076	37.4	5	- 0.14	0.15
80	Halocynthia roretzi	14771	31.7	13	- 0.29	0.46
81	Balanoglossus carnosus	15708	48.6	9	- 0.03	- 0.30
82	Laqueus rubellus	14017	41.6	5	- 0.29	0.27
83	Terebratulina retusa	15451	42.8	5	0.03	- 0.29
84	Arbacia lixula	15719	37.5	9	- 0.06	- 0.09
85	Paracentrotus lividus	15696	39.7	9	0.02	- 0.13
86	Strongylocentrotus purpuratus	15650	41.0	9	- 0.03	- 0.10
87	Asterina pectinifera	16260	38.7	9	0.06	- 0.27
88	Florometra serratissima	16005	27.2	9	- 0.27	0.15
89	Apis mellifera ligustica	16343	15.1	5	0.02	- 0.27
90	Bombyx mori	15643	18.7	5	0.06	- 0.22
91	Ceratitis capitata	15980	22.5	5	0.02	- 0.19
92	Drosophila melanogaster	19517	17.8	5	0.02	- 0.15
93	Drosophila yakuba	16019	21.4	5	0.01	- 0.14
94	Anopheles gambiae	15363	22.4	5	0.03	- 0.15
95	Anopheles quadrimaculatus	15455	22.6	5	0.04	- 0.18
96	Locusta migratoria	15722	24.7	5	0.18	- 0.18
97	Artemia franciscana	15822	35.6	5	- 0.04	- 0.01
98	Daphnia pulex	15333	37.7	5	0.01	- 0.12
99	Penaeus monodon	15984	29.4	5	0	- 0.14
100	Ixodes hexagonus	14539	27.3	5	0.03	- 0.37
101	Rhipicephalus sanguineus	14710	22.0	5	- 0.03	- 0.1
102	Lumbricus terrestris	14998	38.4	5	- 0.03	- 0.18
103	Platynereis dumerilii	15619	35.9	5	- 0.03	- 0.14
104	Albinaria coerulea	14130	29.4	5	- 0.07	0.06

105	Cepaea nemoralis	14100	40.2	5	- 0.12	0.06
106	Pupa strigosa	14189	38.9	5	- 0.1	0.06
107	Crassostrea gigas	18224	36.7	5	- 0.13	0.2
108	Katharina tunicata	15532	30.5	5	- 0.10	0.22
109	Loligo bleekeri	17211	28.7	5	0.09	- 0.36
110	Ascaris suum	14284	28.0	5	- 0.38	0.45
111	Caenorhabditis elegans	13794	23.8	5	- 0.18	0.25
112	Onchocerca volvulus	13747	26.7	5	- 0.47	0.49
113	Fasciola hepatica	14462	37.8	5	- 0.48	0.47
114	Paragonimus westermani	14964	48.3	14	- 0.39	0.29
115	Schistosoma japonicum	14085	29.0	5	- 0.30	0.42
116	Schistosoma mansoni	14415	31.5	5	- 0.26	0.46
117	Schistosoma mekongi	14072	27.8	5	- 0.28	0.48
118	Echinococcus multilocularis	13738	31.0	14	- 0.40	0.51
119	Taenia crassiceps	13503	26.0	5	- 0.31	0.41
120	Metridium senile	17443	38.1	4	- 0.13	0.11
121	Pichia canadensis	27694	18.1	4	0.02	0.12
122	Saccharomyces cerevisiae	85779	17.1	3	0.02	0.06
123	Podospora anserina	100314	30.1	4	0.02	0.11
124	Schizosaccharomyces pombe	19431	30.1	4	- 0.03	0.05
125	Allomyces macrogynus	57473	39.5	4	- 0.04	0.06
126	Arabidopsis thaliana	366923	44.8	1	0.01	- 0.01
127	Beta vulgaris var. altissima	368799	43.9	1	0	0
128	Marchantia polymorpha	186609	42.4	1	- 0.01	0.01
129	Chlamydomonas eugametos	22897	34.6	1	0.01	0.15
130	Chlamydomonas reinhardtii	15758	45.2	1	0.01	0.01
131	Scenedesmus obliquus	42781	36.2	22	- 0.04	0.12
132	Pedinomonas minor	25137	22.2	4	- 0.19	0.13
133	Prototheca wickerhamii	55328	25.8	1	0.03	- 0.02
134	Rhodomonas salina	48063	29.8	1	0.05	0.06
135	Chondrus crispus	25836	27.9	4	0.05	0.04
136	Cyanidioschyzon merolae	32211	27.1	1	0.01	0.06
137	Porphyra purpurea	36753	33.5	4	0.07	0.04
138	Paramecium aurelia	40469	41.2	4	0.14	0.06
139	Plasmodium falciparum	5967	31.6	4	- 0.05	0.01
140	Plasmodium reichenowi	5966	31.7	4	- 0.05	0.01

141	Tetrahymena pyriformis	47296	21.3	4	- 0.04	0.02
142	Acanthamoeba castellanii	41591	29.4	4	- 0.09	0.11
143	Naegleria gruberi	49843	22.2	1	- 0.09	0.17
144	Leishmania tarentolae	20992	21.1	1	- 0.03	0.10
145	Physarum polycephalum	65862	25.9	1	0.03	0.02
146	Dictyostelium discoideum	55564	27.4	1	0.20	0.24
147	Malawimonas jakobiformis	47328	26.1	1	0.03	0
148	Reclinomonas americana	69034	26.1	1	0	0.13
149	Cafeteria roenbergensis	43159	27.3	4	- 0.05	0.11
150	Chrysodidymus synuroideus	34119	24.1	1	0.02	0.05
151	Ochromonas danica	41035	26.2	1	0.06	0.01
152	Phytophthora infestans	37957	22.3	1	0	0.05

配列は９つのグループに分けられる；哺乳類（１～４０）、鳥類（４１～５０）、爬虫類（５１～５６）、両生類（５７、５８）、魚類（５９～８０）(Cephalochordata and Urochordata を含む)、無脊椎動物（８１～１２０）(Hemichordata を含む)、菌類（１２１～１２５）、植物（１２６～１３７）、原生生物（１３８～１５２）。

[a] GC 含量（%）

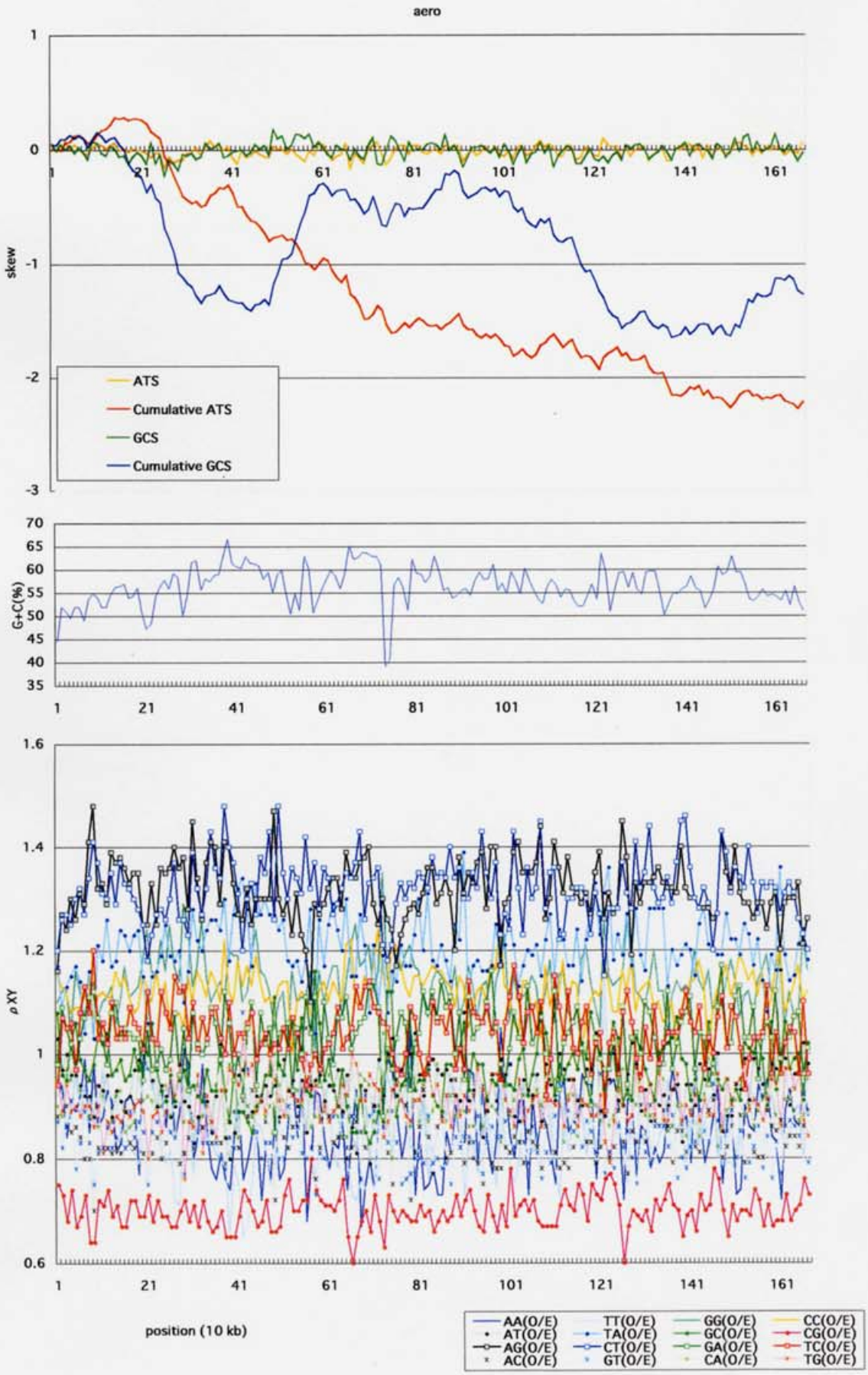
[b] 遺伝暗号の分類番号（NCBI のデータより）

[c] A と T（ATS）、G と C（GCS)間の塩基組成の偏り。

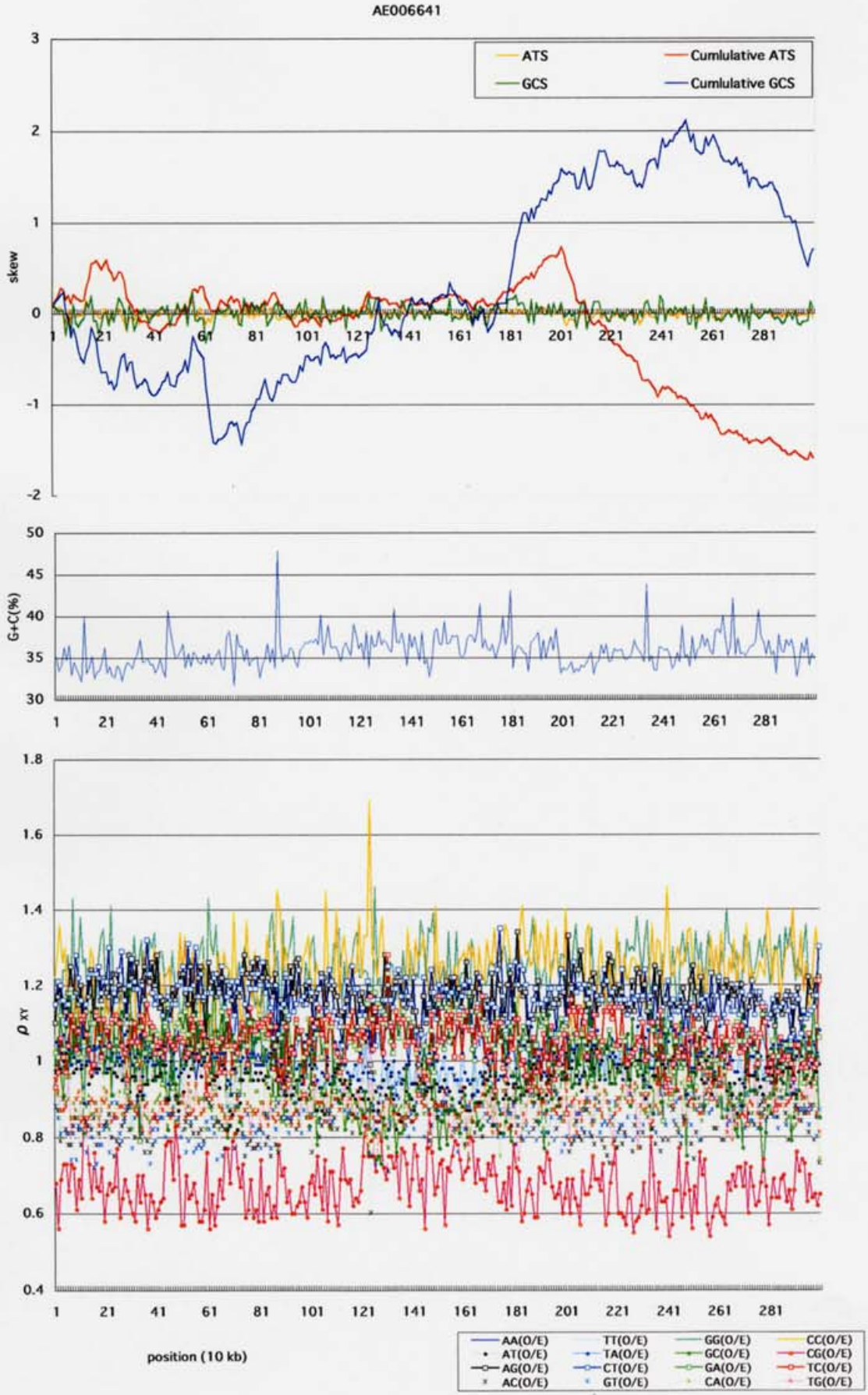
付録 B　本文中に関連する図

Archaea; Crenarchaeota (3)

A-1 *Aeropyrum pernix* NC_000854 1,669,695 11 56.3

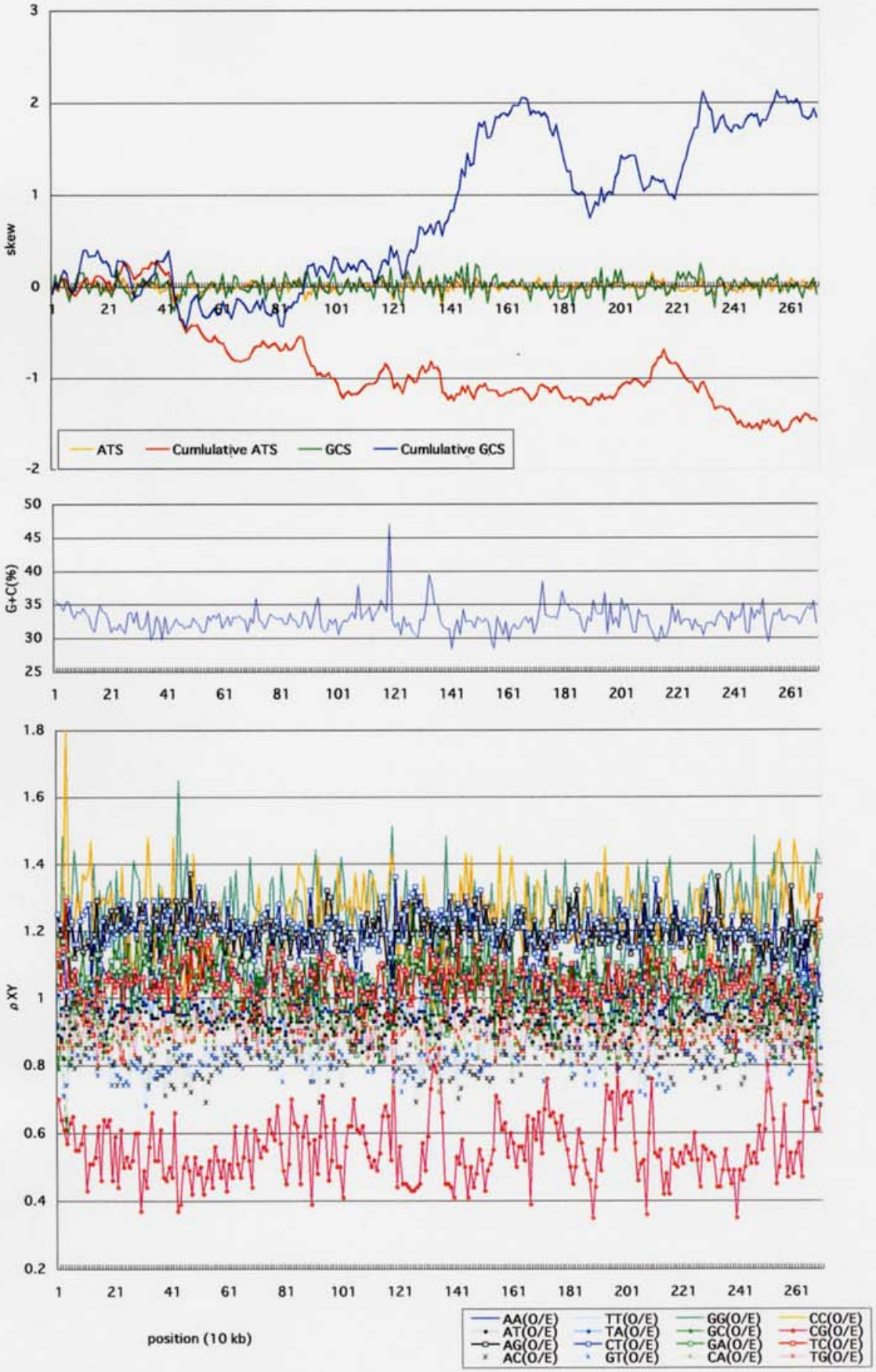


A-8 *Sulfolobus solfataricus* NC_002754 2,992,245 11 35.8



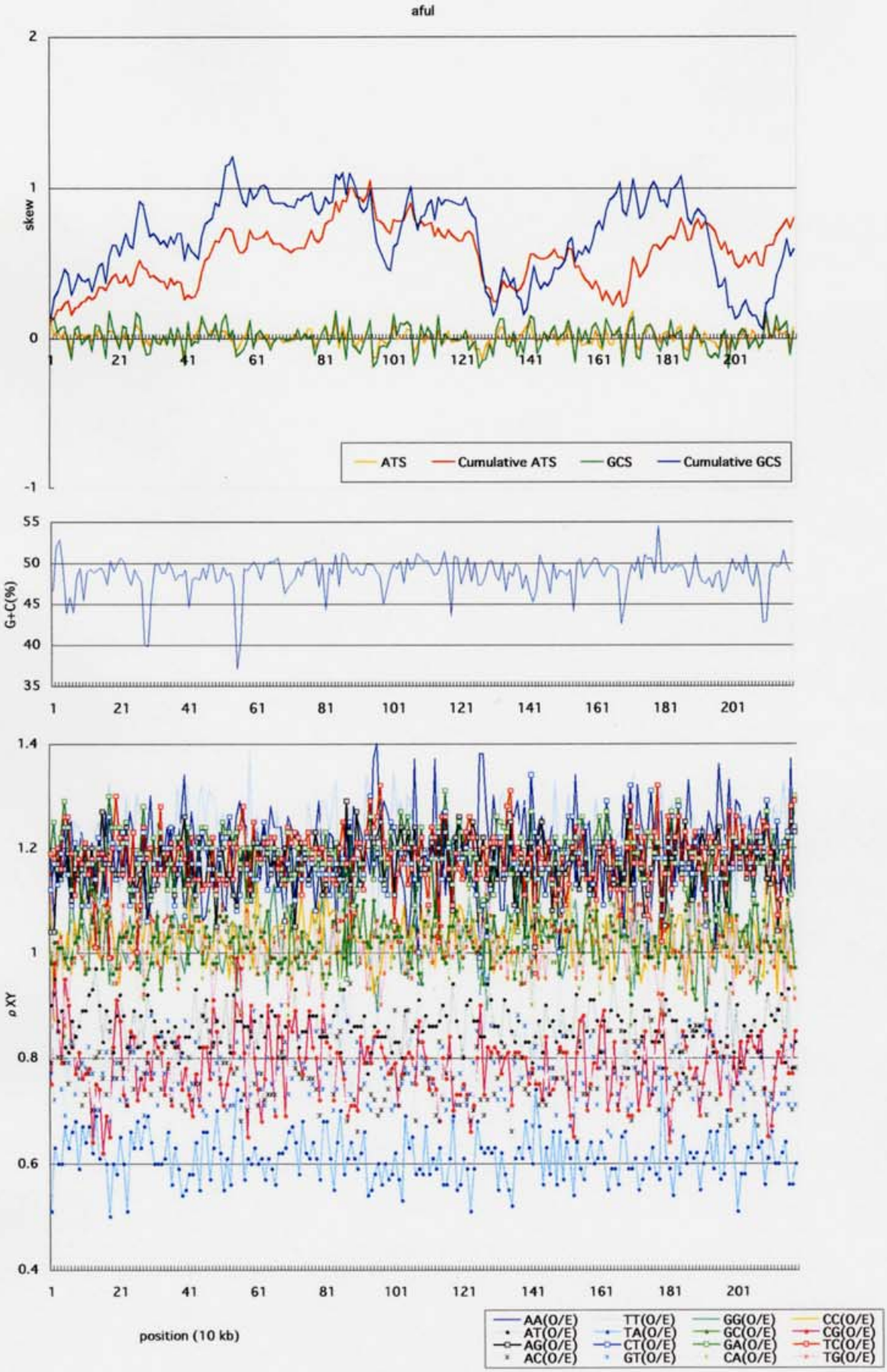
A-9 *Sulfolobus tokodaii* NC_003106 2,694,765 11 32.8

BA000023

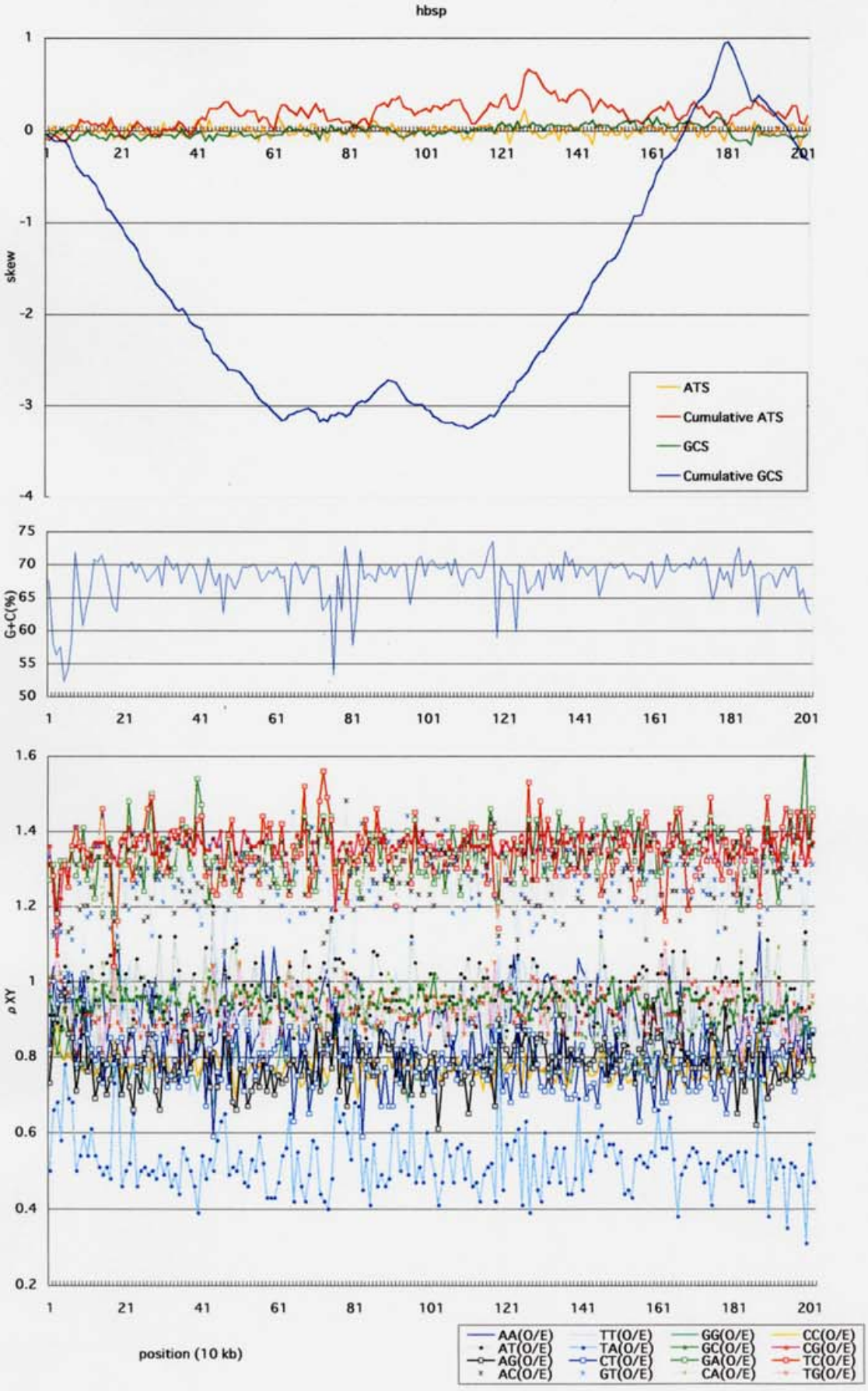


Archaea; Euryarchaeota (8)

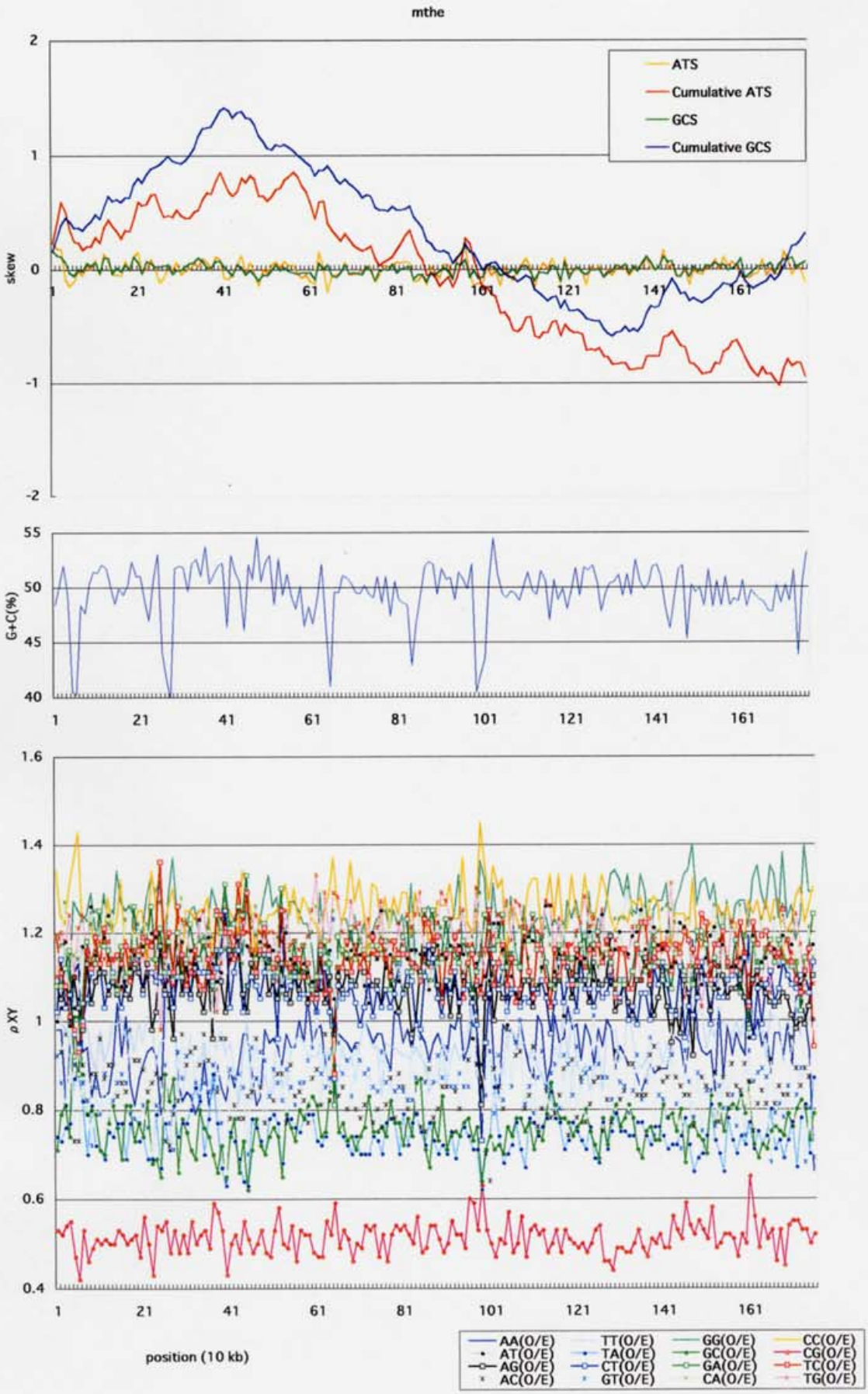
A-2 *Archaeoglobus fulgidus* NC_000917 2,178,400 11 48.6



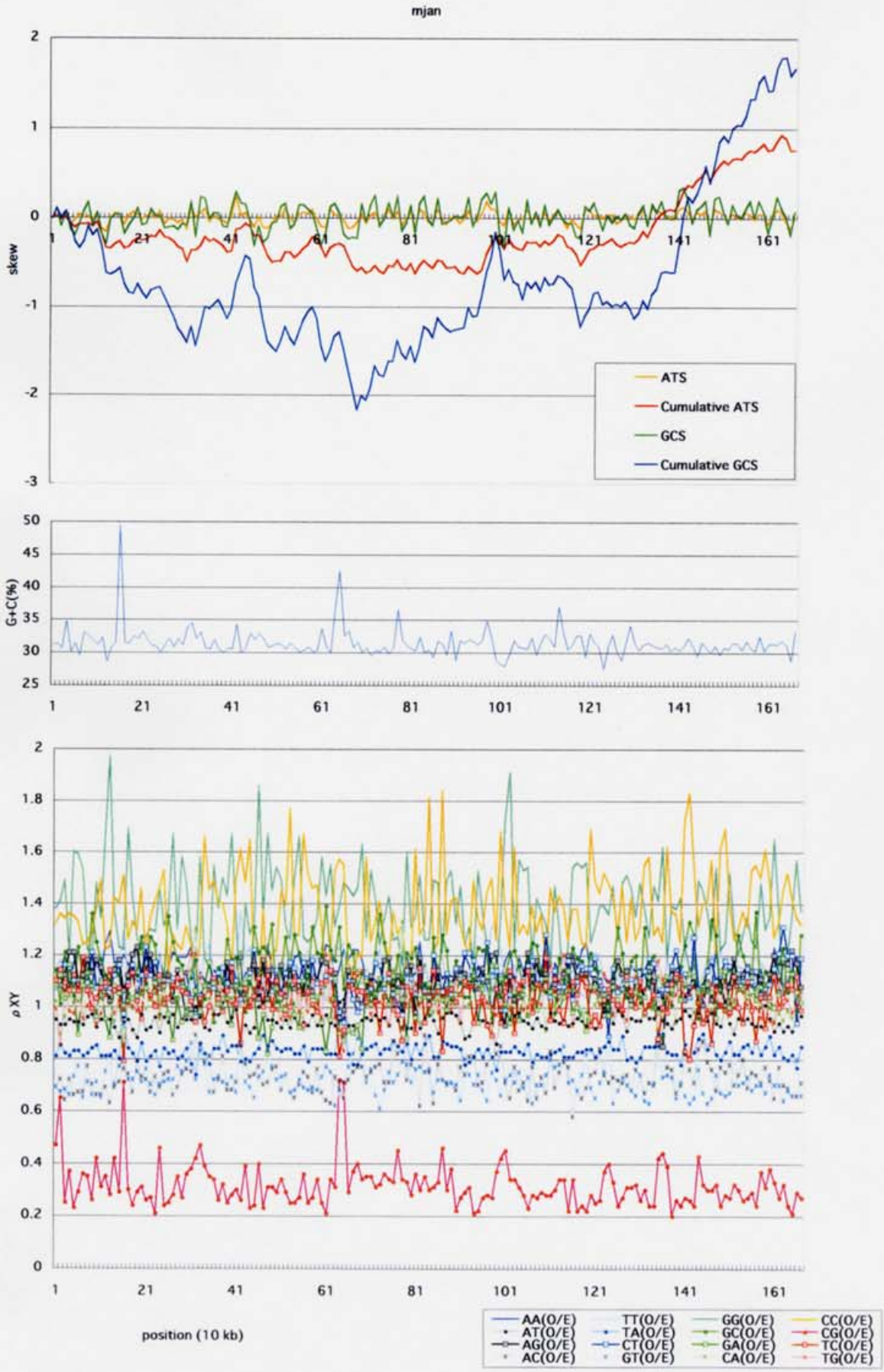
A-3 *Halobacterium sp. NRC-1* NC_002607 2,014,239 11 67.9



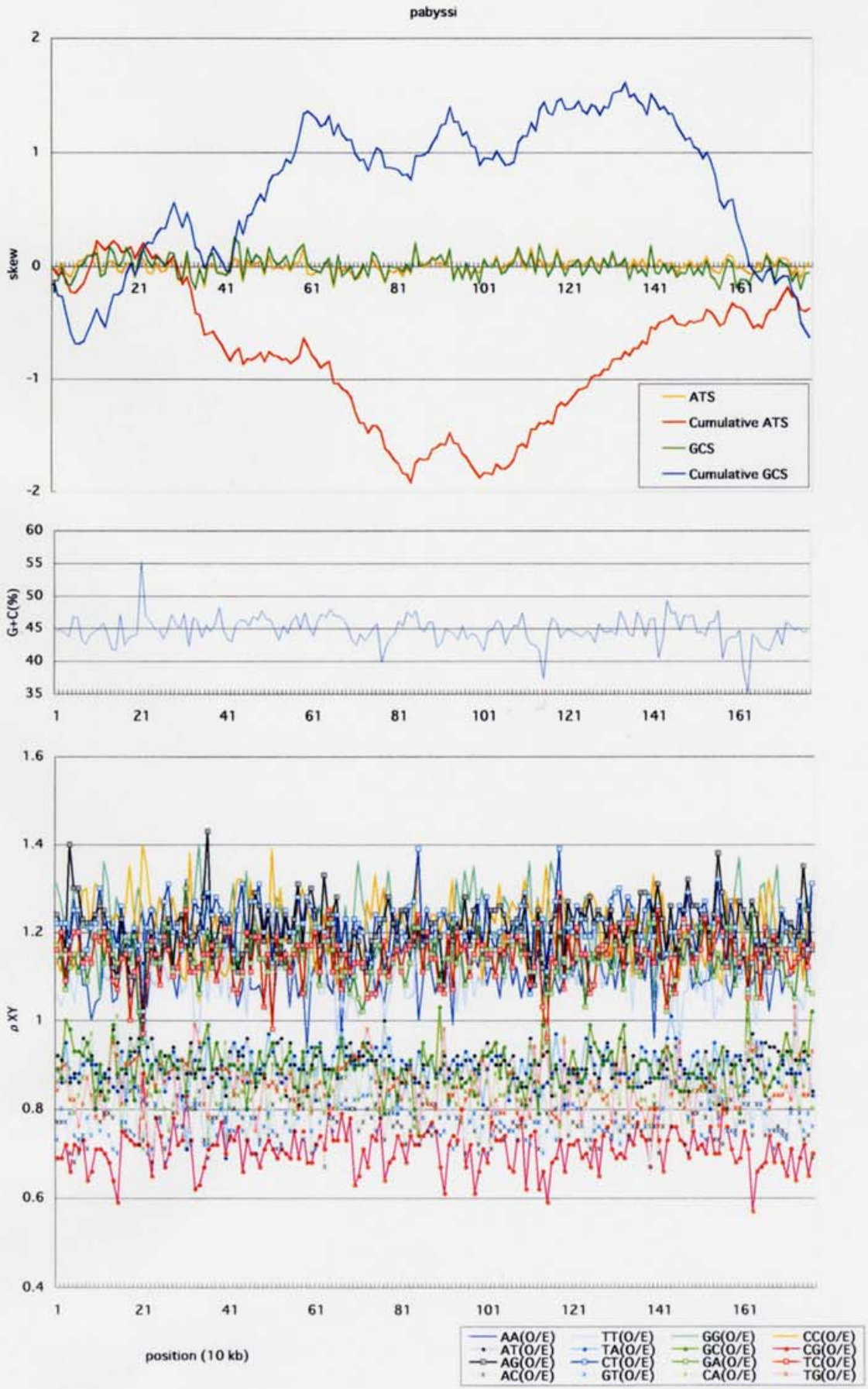
A-4 *Methanobacterium thermoautotrophicum* NC_000916 1,751,377 11 49.5



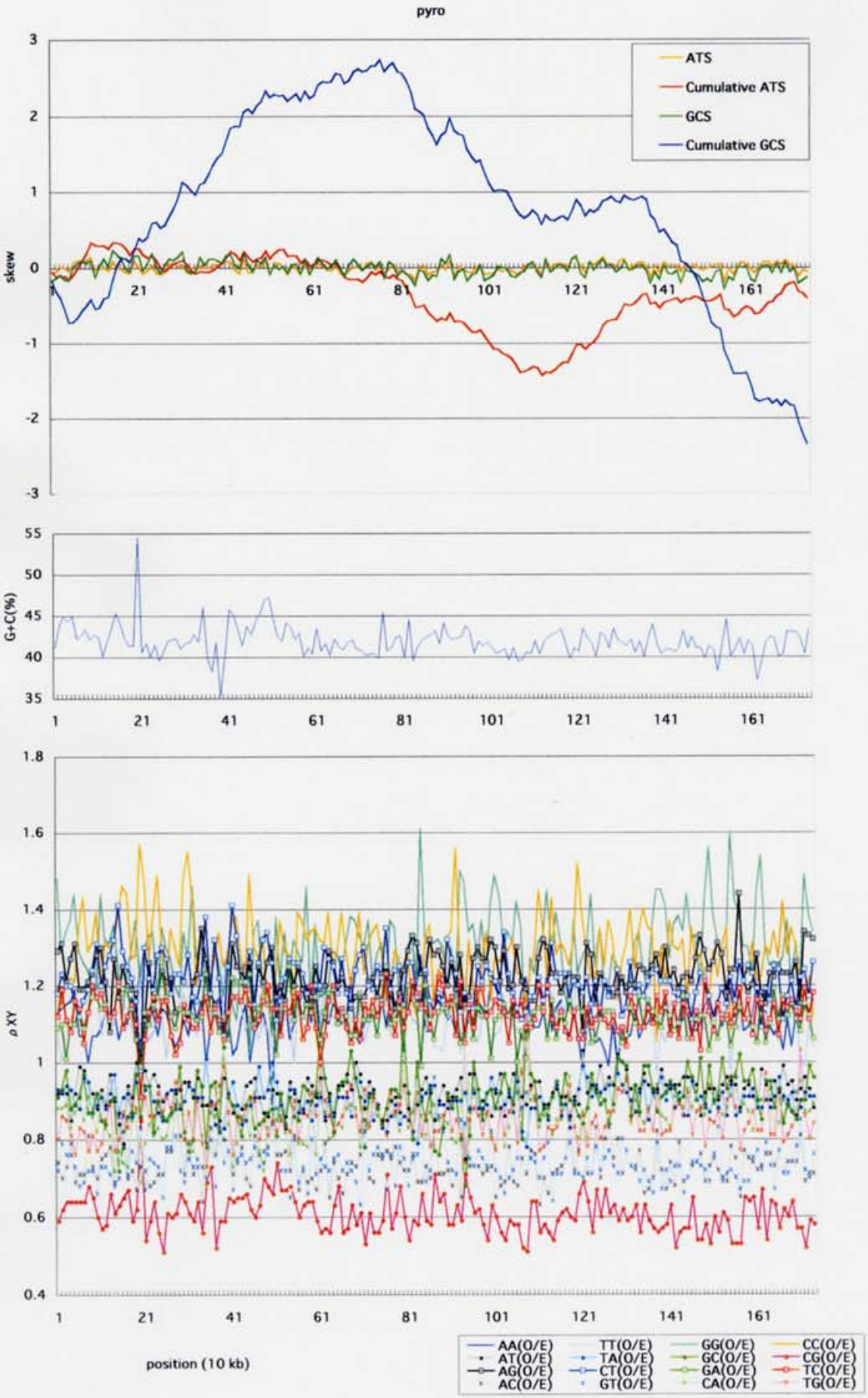
A-5 *Methanococcus jannaschii* NC_000909 1,664,970 11 31.4



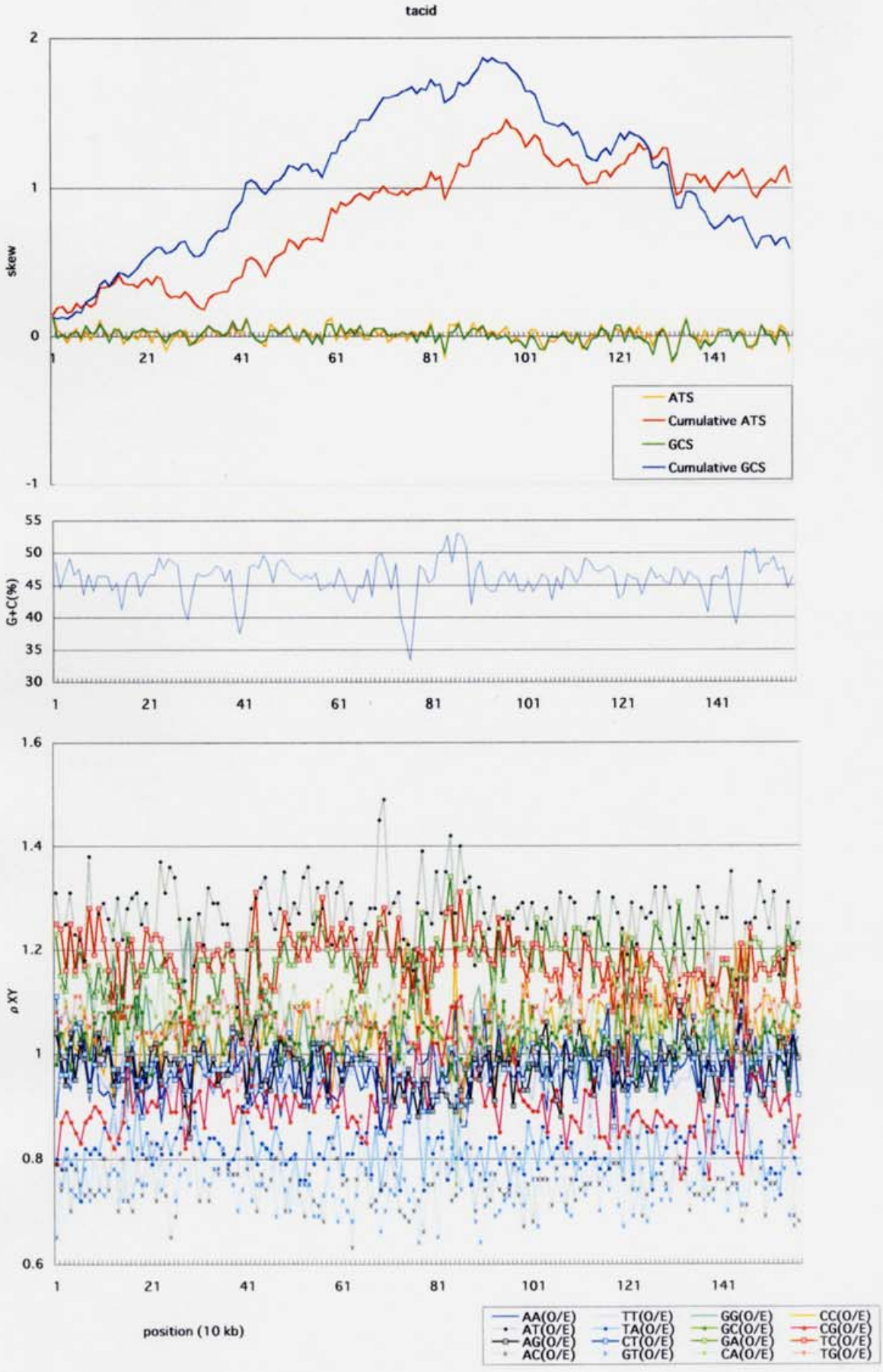
A-6 *Pyrococcus abyssi* NC_000868 1,765,118 11 44.7



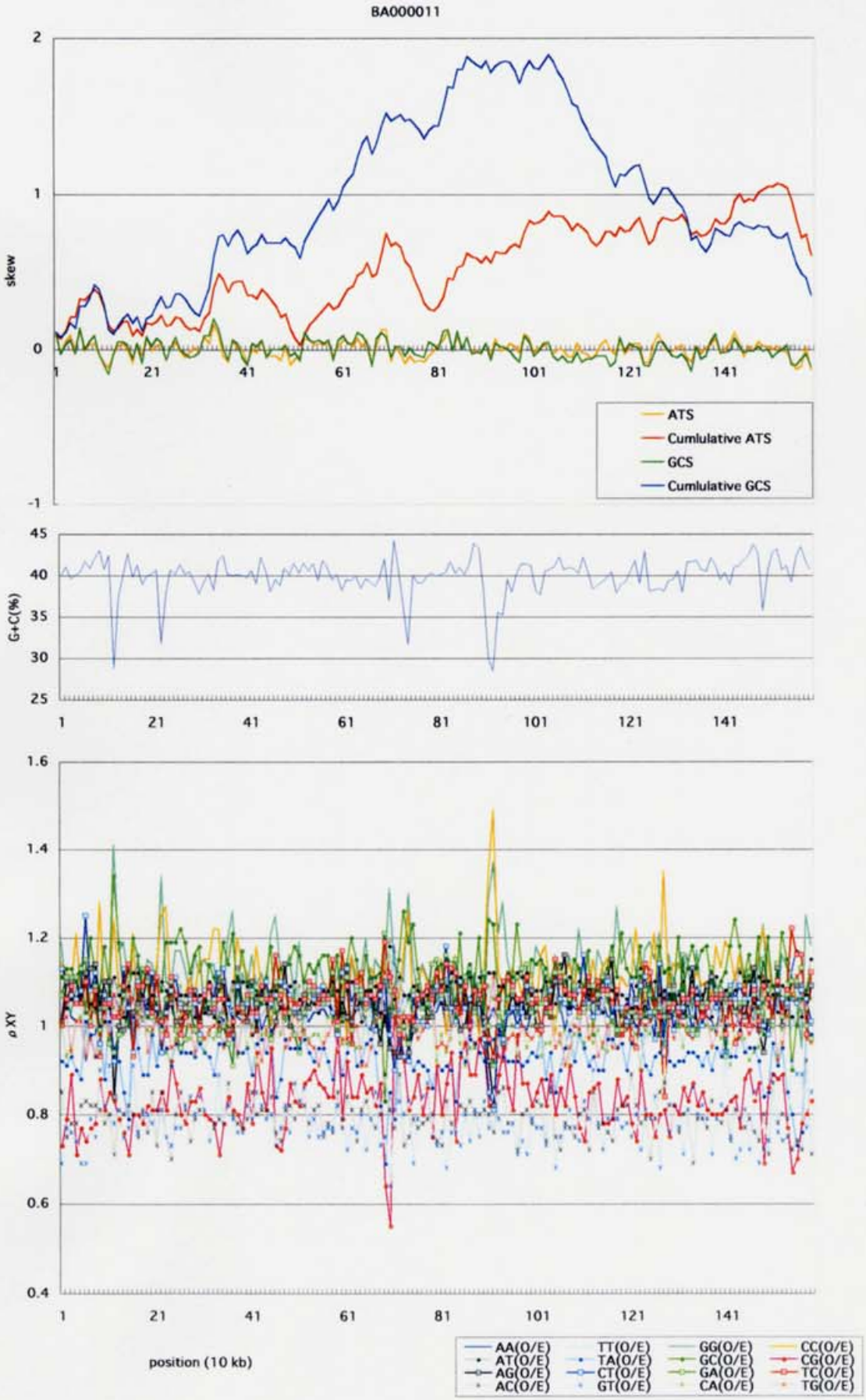
A-7 *Pyrococcus horikoshii* NC_000961 1,738,505 11 41.9



A-10 *Thermoplasma acidophilum* NC_002578 1,564,906 11 46.0



A-11 *Thermoplasma volcanium* NC_002689 1,584,804 11 39.9

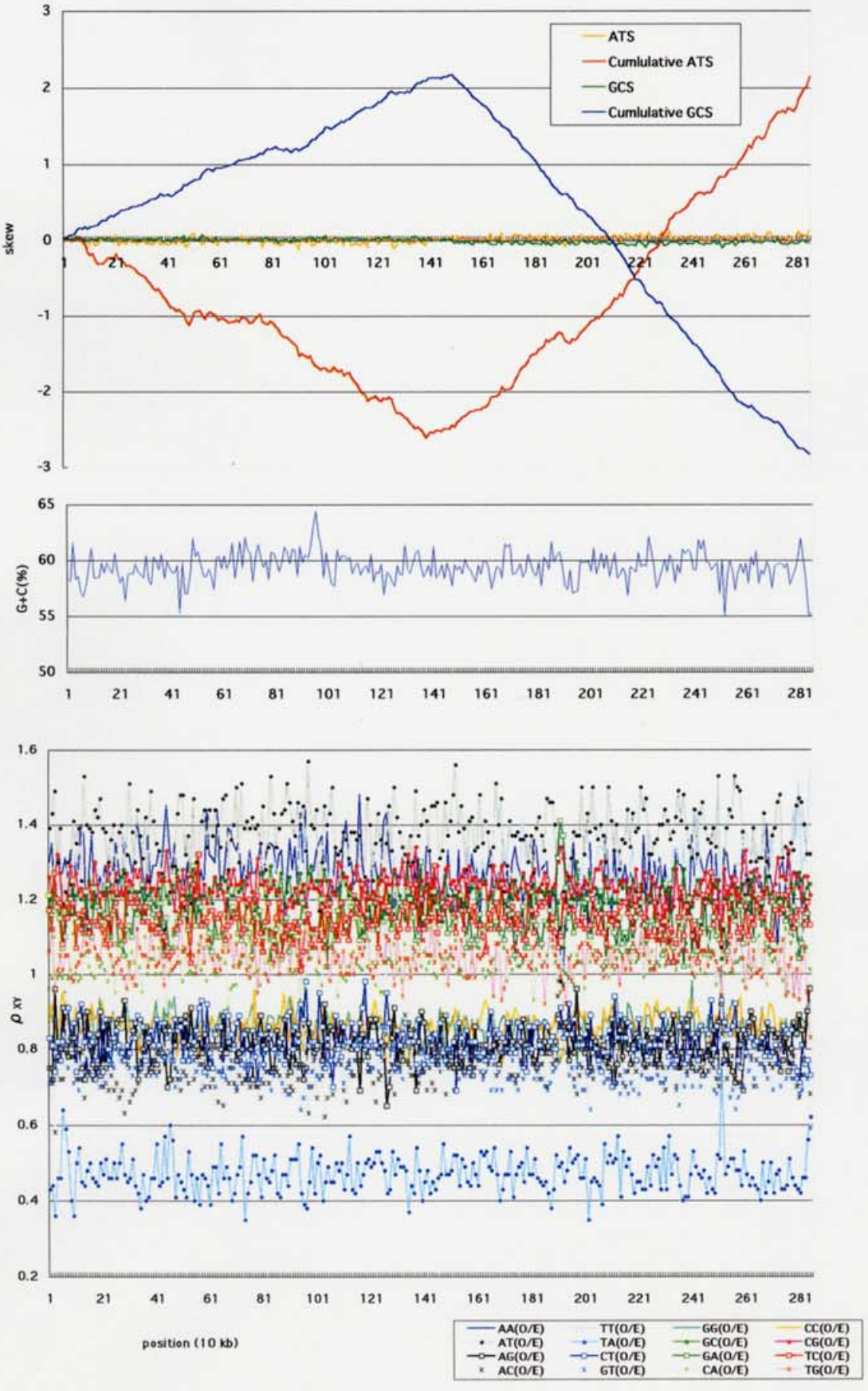


Bacteria; Proteobacteria;

alpha subdivision (7)

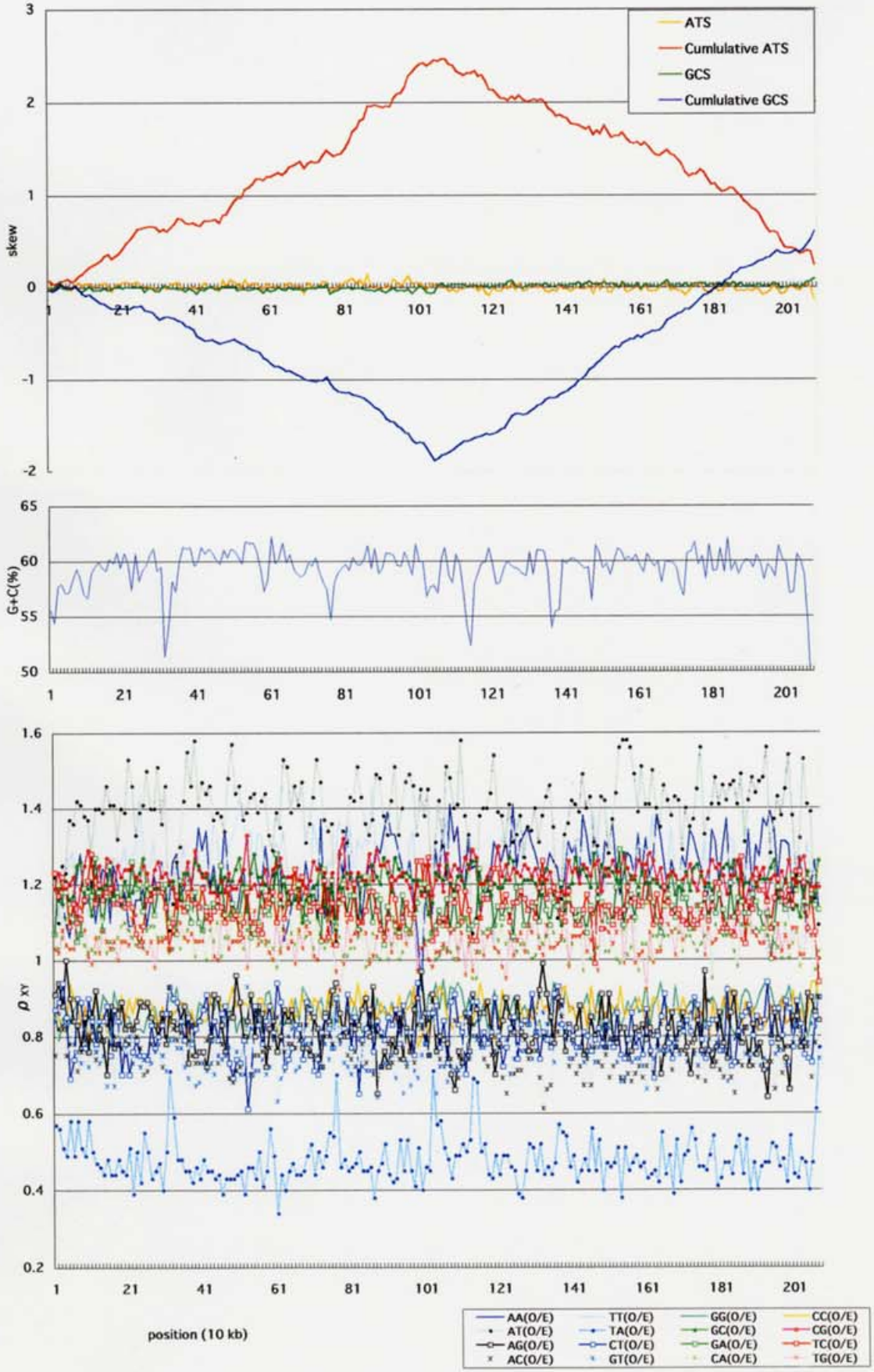
B-1 *Agrobacterium tumefaciens* circular NC_003062 2,841,581 11 59.4

AE007869

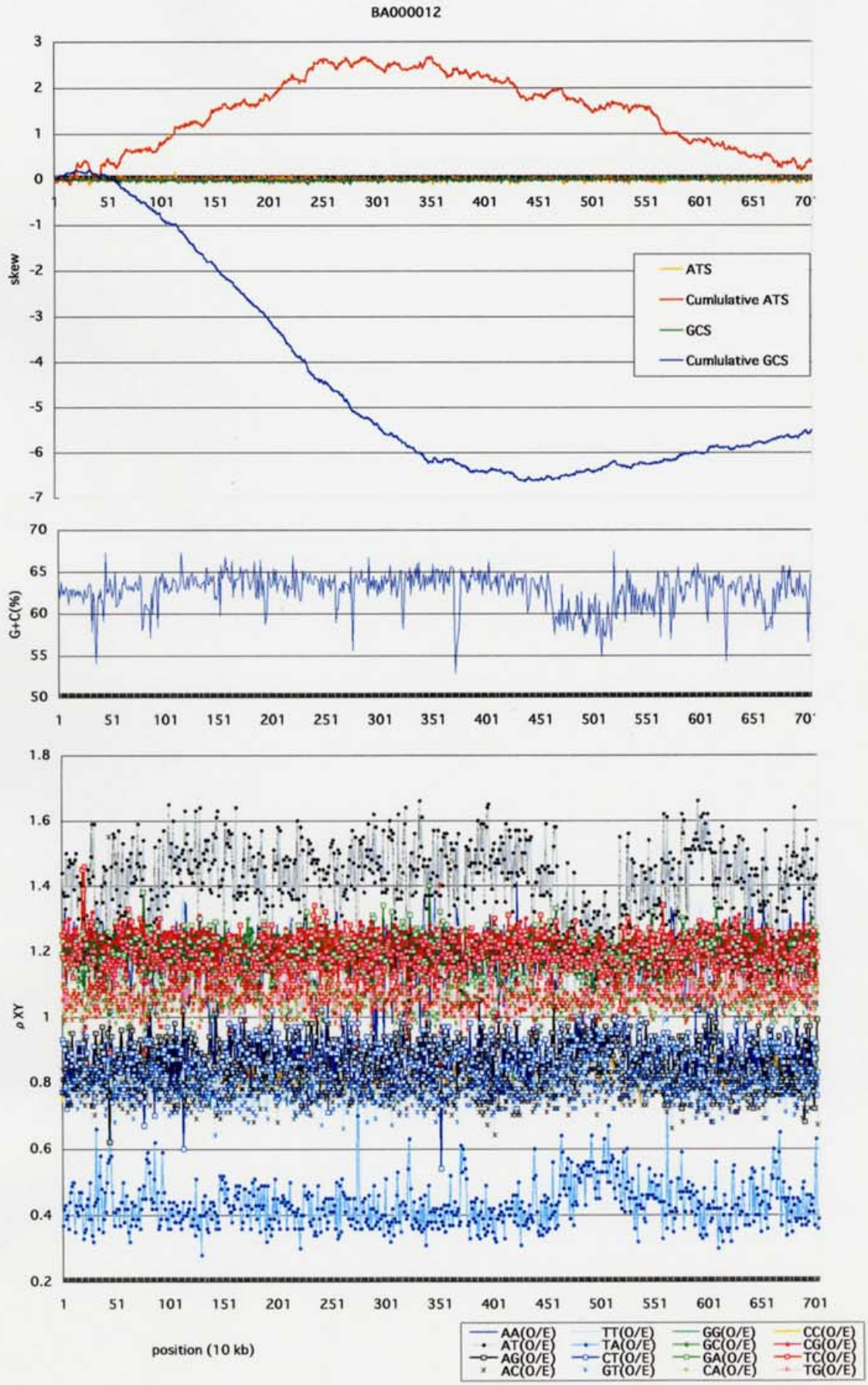


B-2 *Agrobacterium tumefaciens* linear NC_003063 2,074,782 11 59.3

AE007870

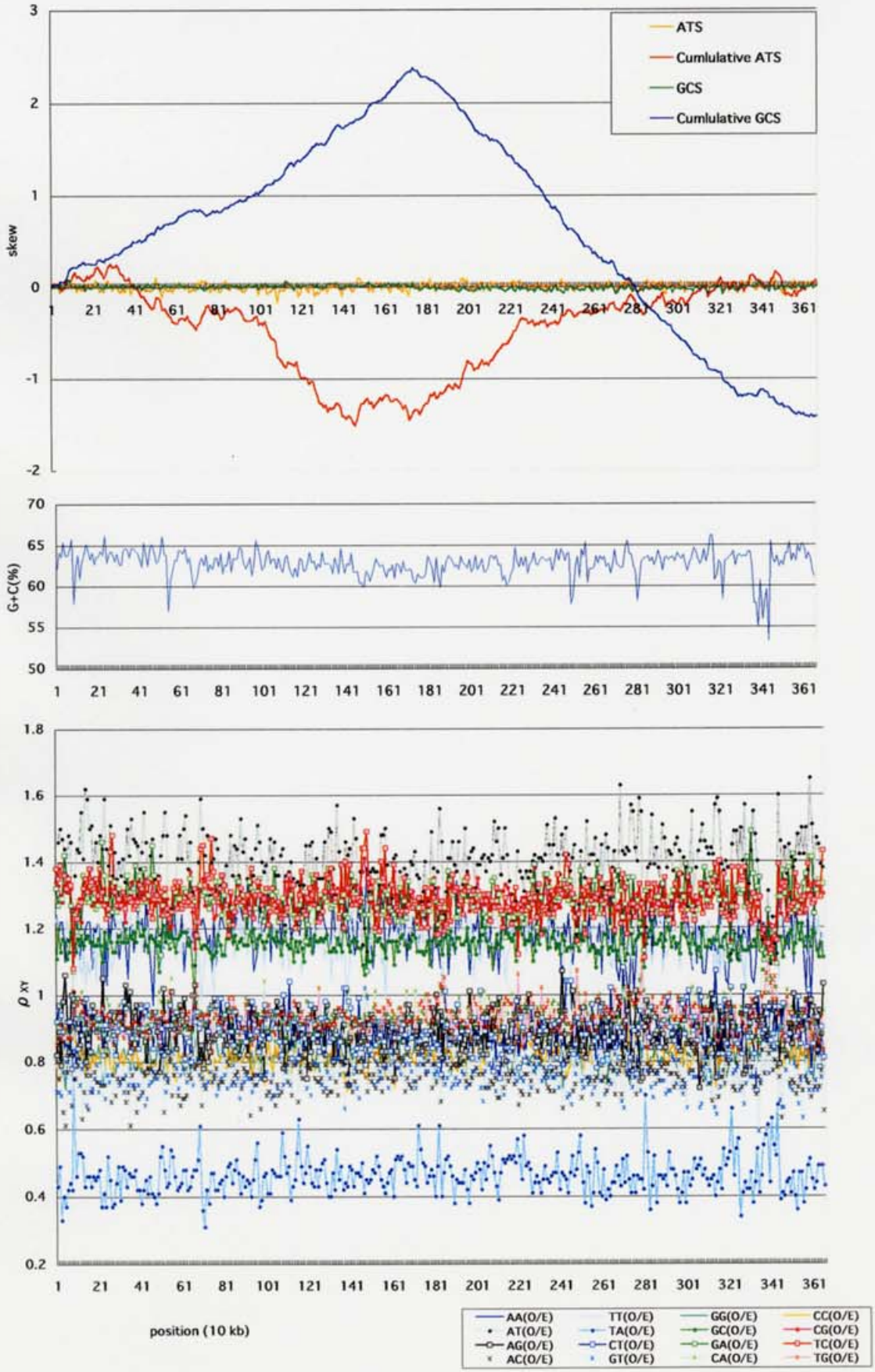


B-27 *Mesorhizobium loti* NC_002678 7,036,074 11 62.8



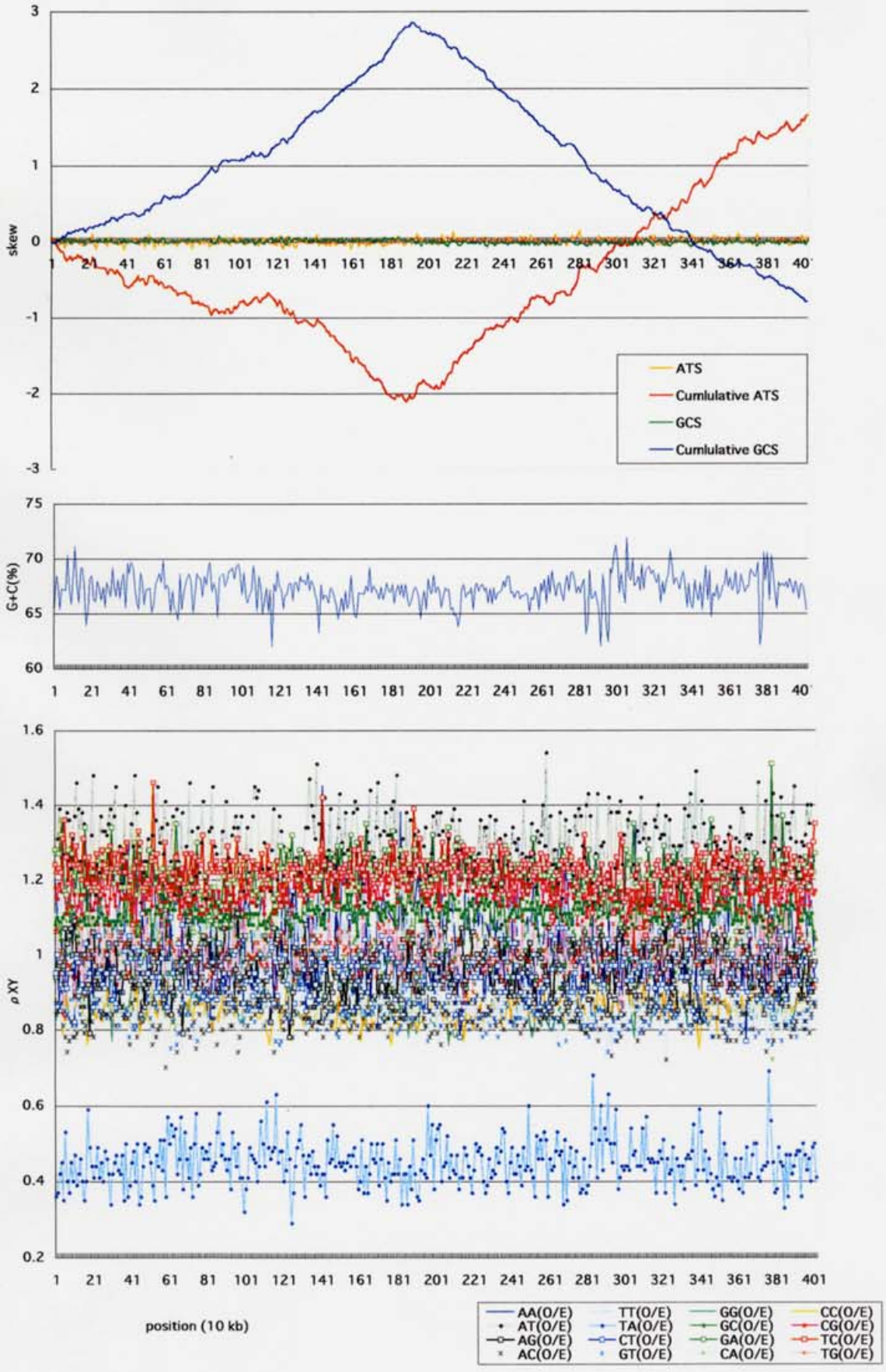
B-42 *Sinorhizobium meliloti* NC_003047 3,654,135 11 62.7

AL591688



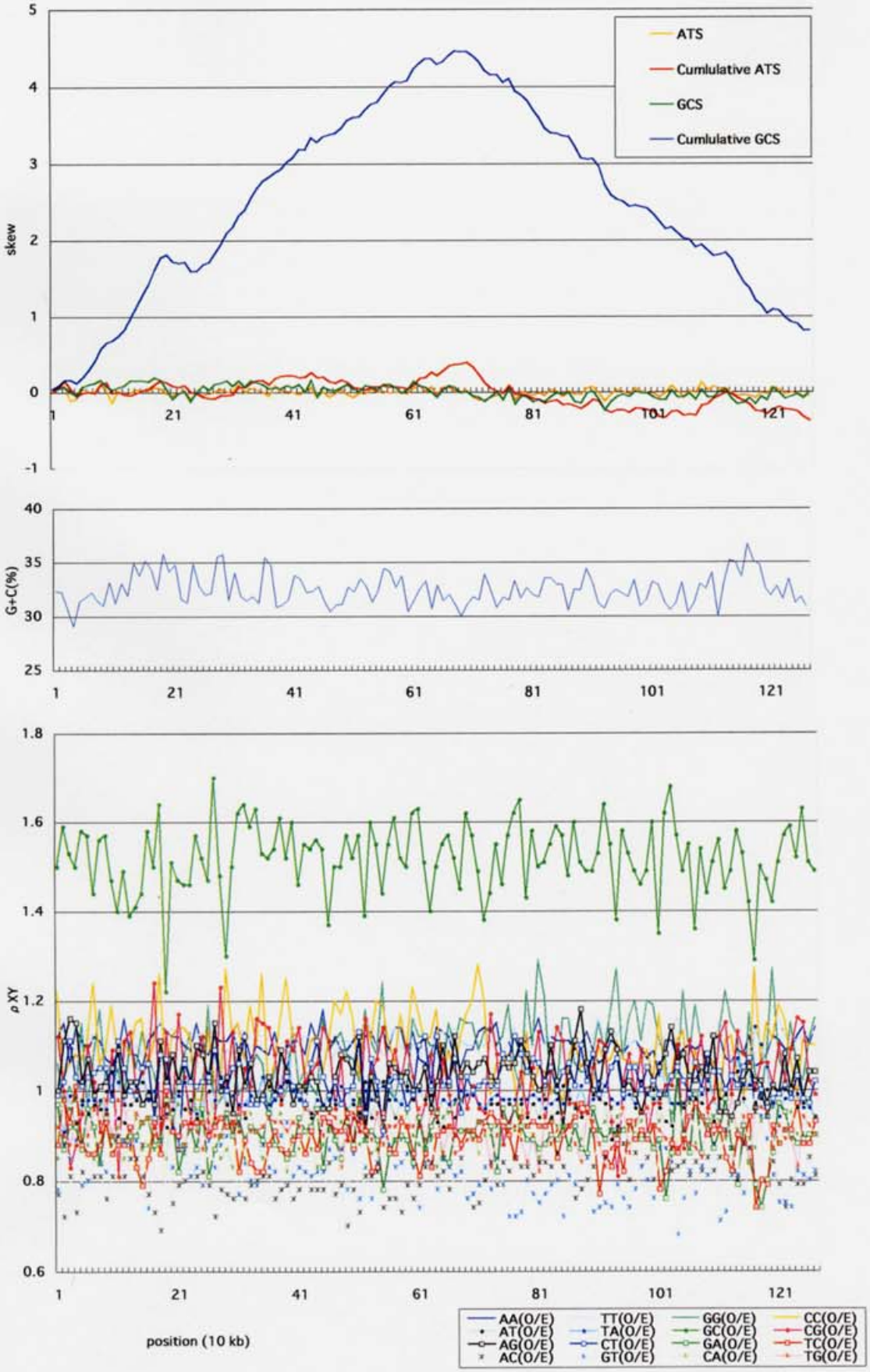
B-9 *Caulobacter crescentus* NC_002696 4,016,947 11 67.2

AE005673

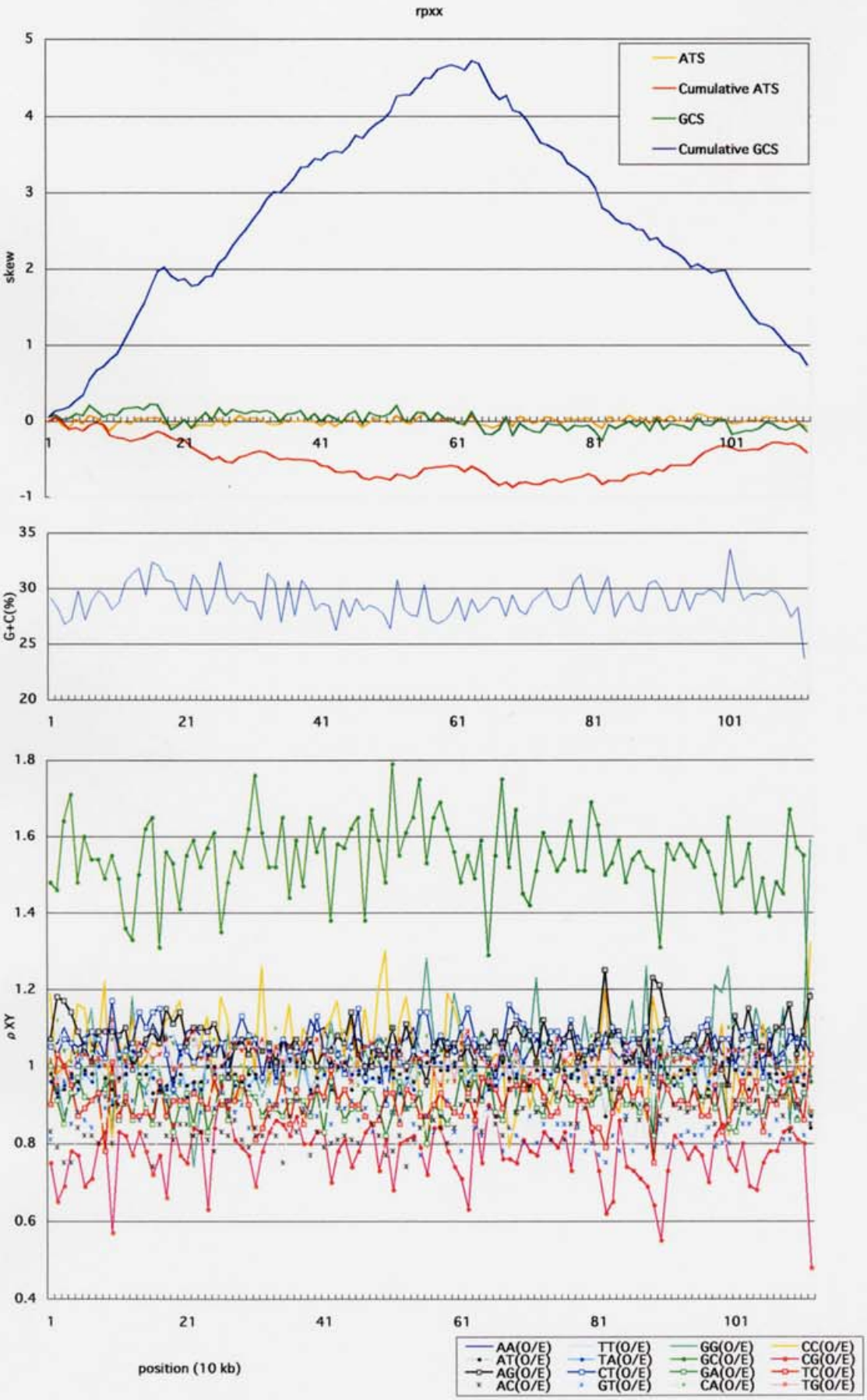


B-38 *Rickettsia conorii* NC_003103 1,268,755 11 32.4

AE006914



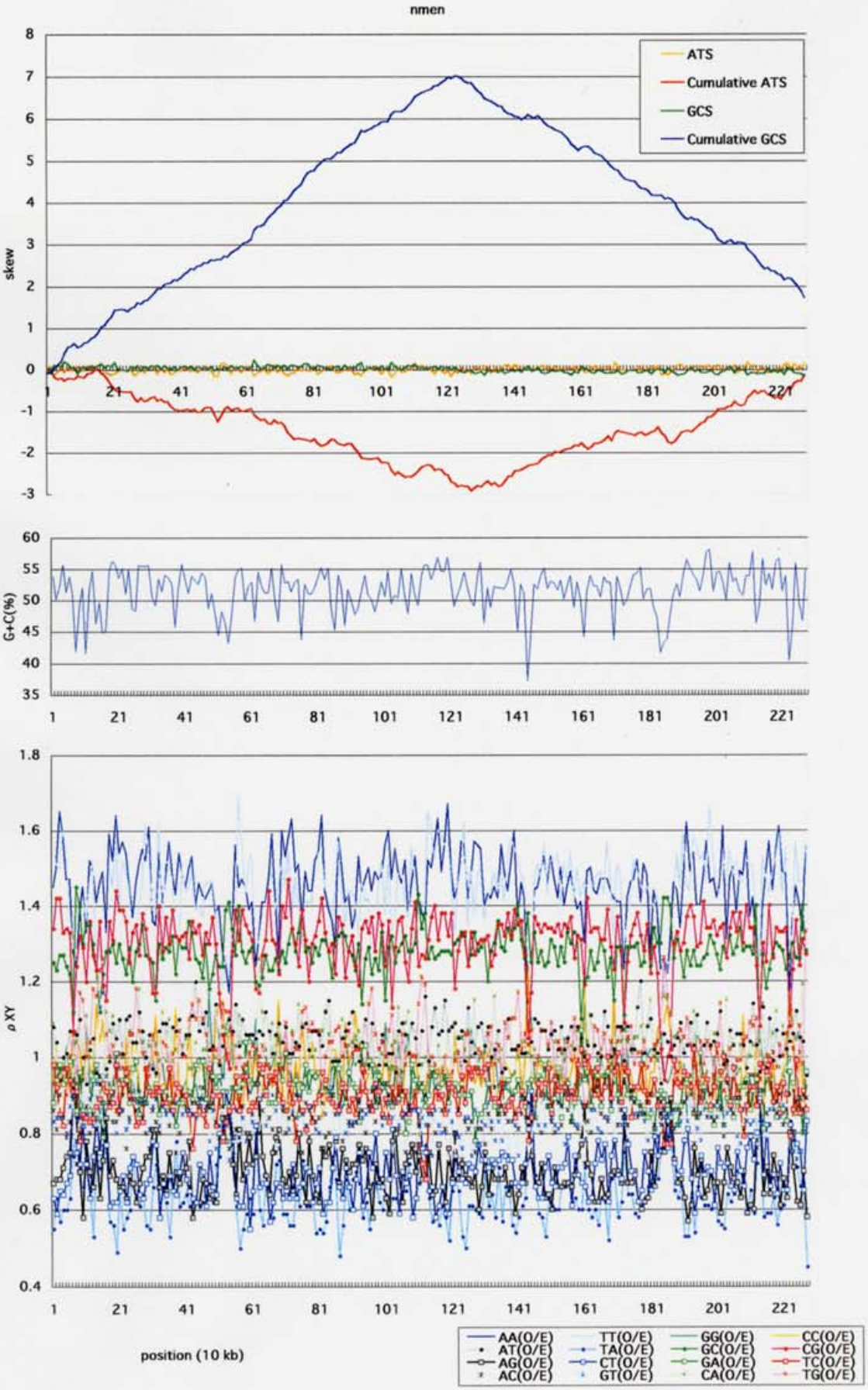
B-39 *Rickettsia prowazekii* NC_000963 1,111,523 11 29.0



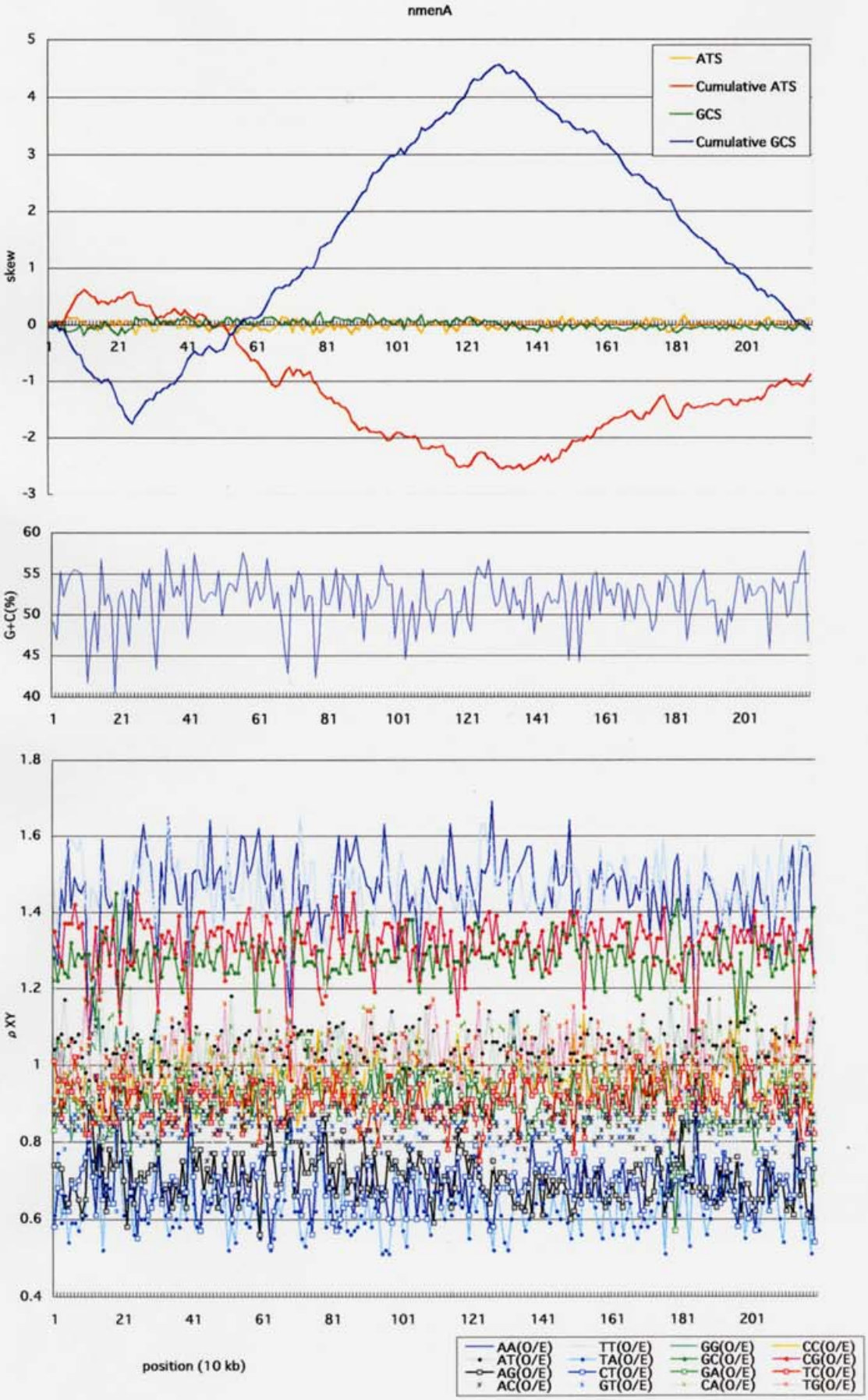
Bacteria; Proteobacteria;

beta subdivision (2)

B-34 *Neisseria meningitidis MC58* NC_003112 2,272,351 11 51.5

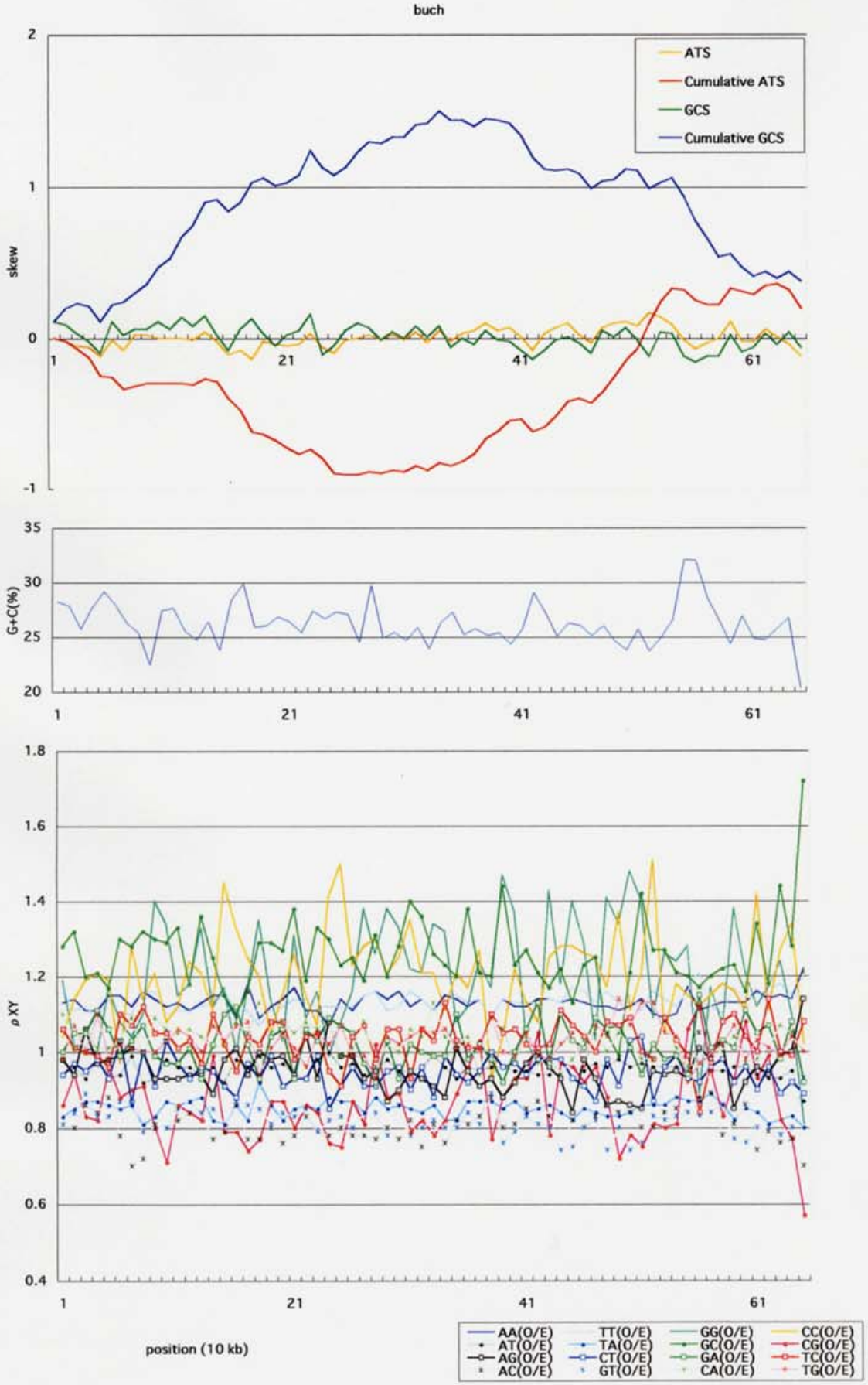


B-35 *Neisseria meningitidis Z2491* NC_003116 2,184,406 11 51.8

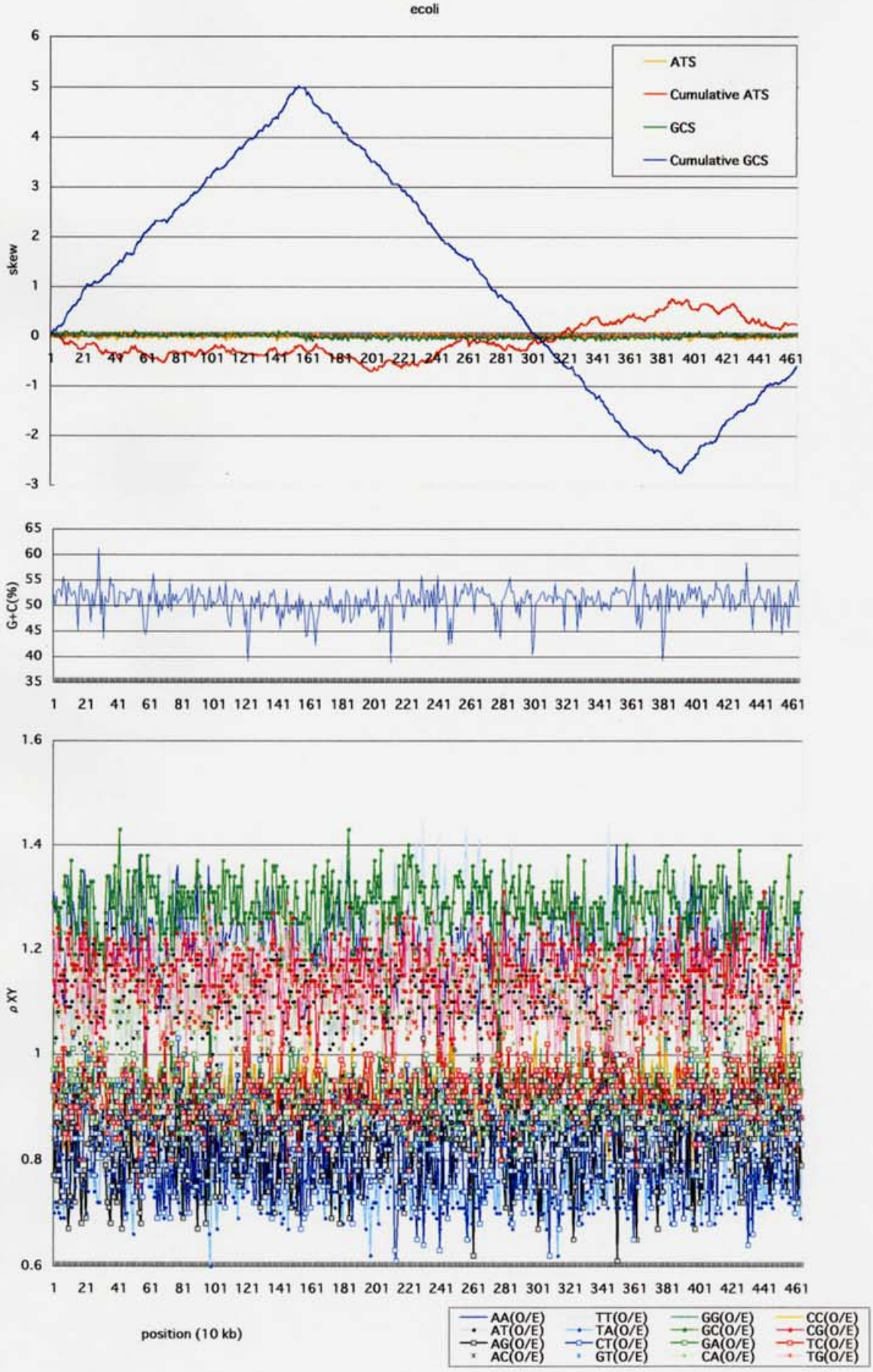


Bacteria; Proteobacteria; gamma subdivision (13)

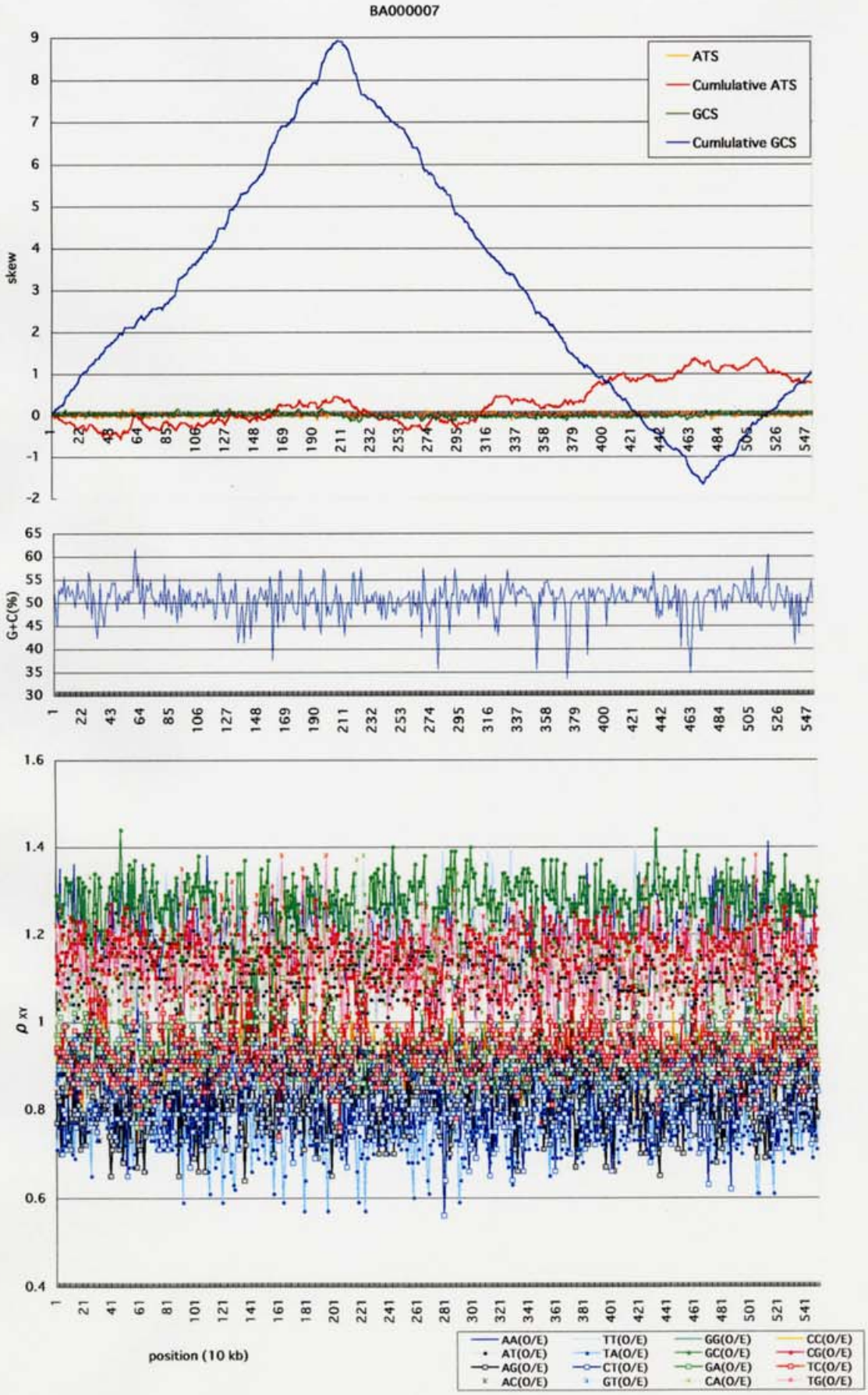
B—7 *Buchnera sp. APS* NC_002528 640,681 11 26.3



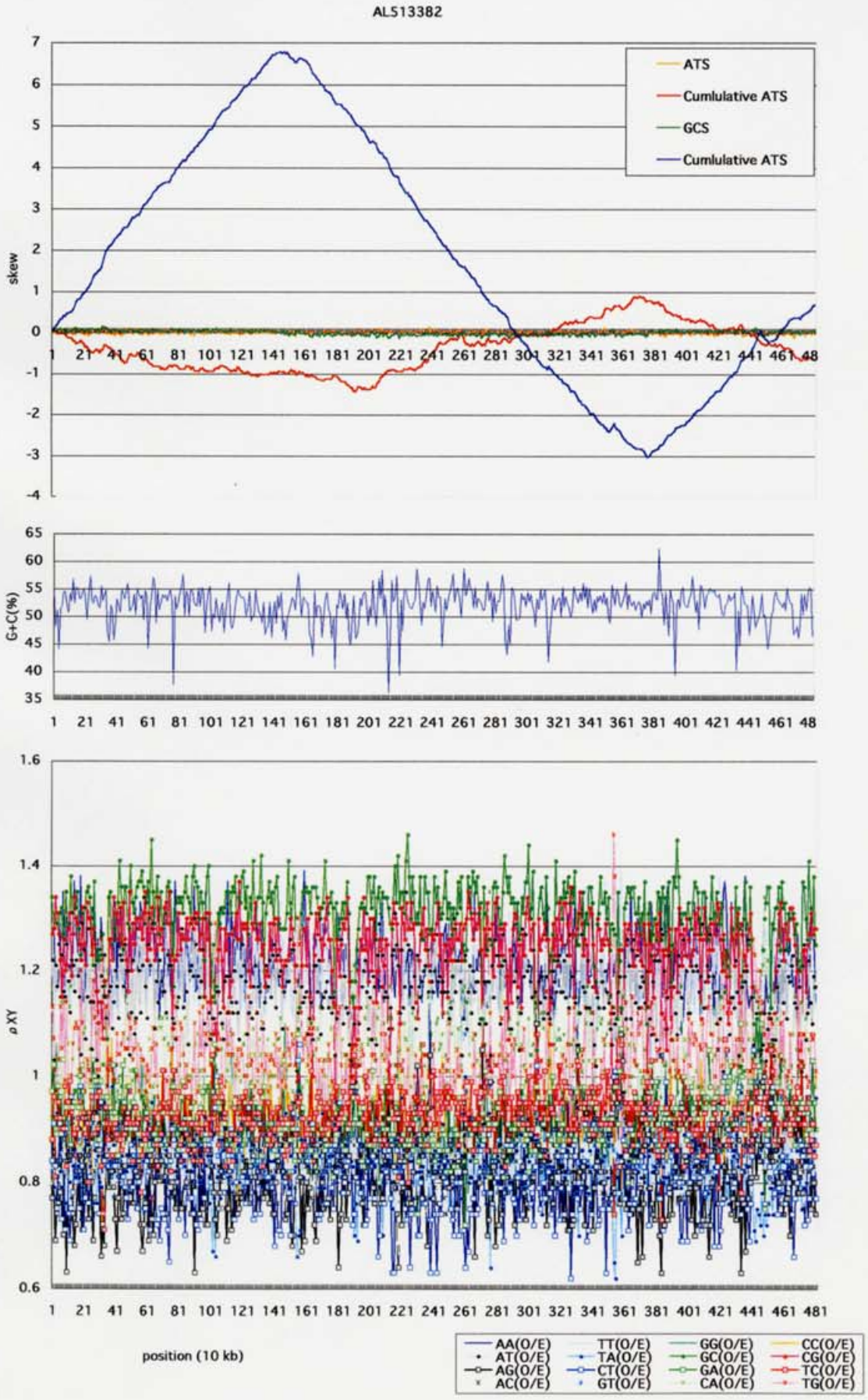
B-18 *Escherichia coli K12* NC_000913 4,639,221 11 50.8



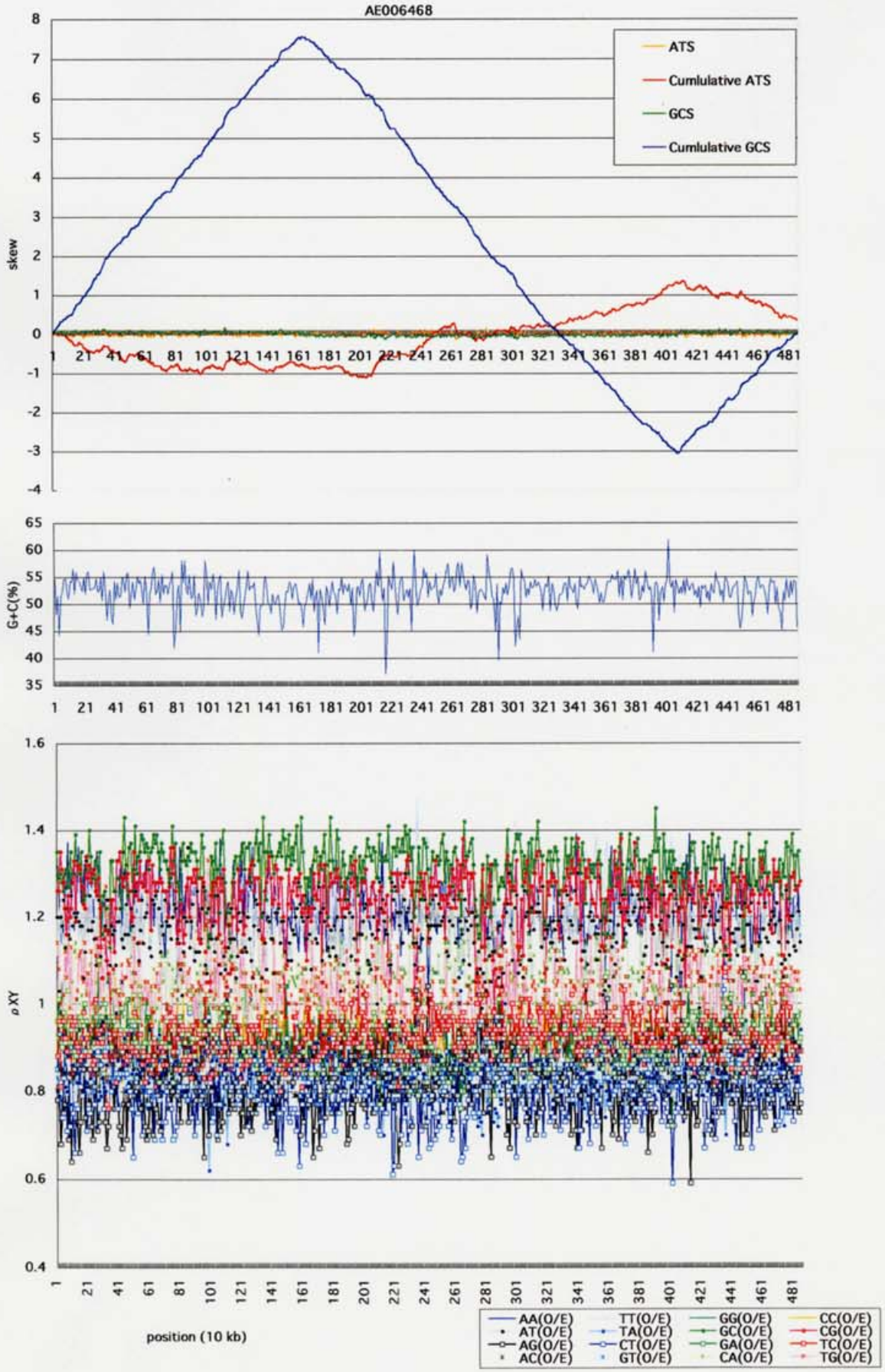
B-19 *Escherichia coli O157:H7* NC_002695 5,498,450 11 50.5



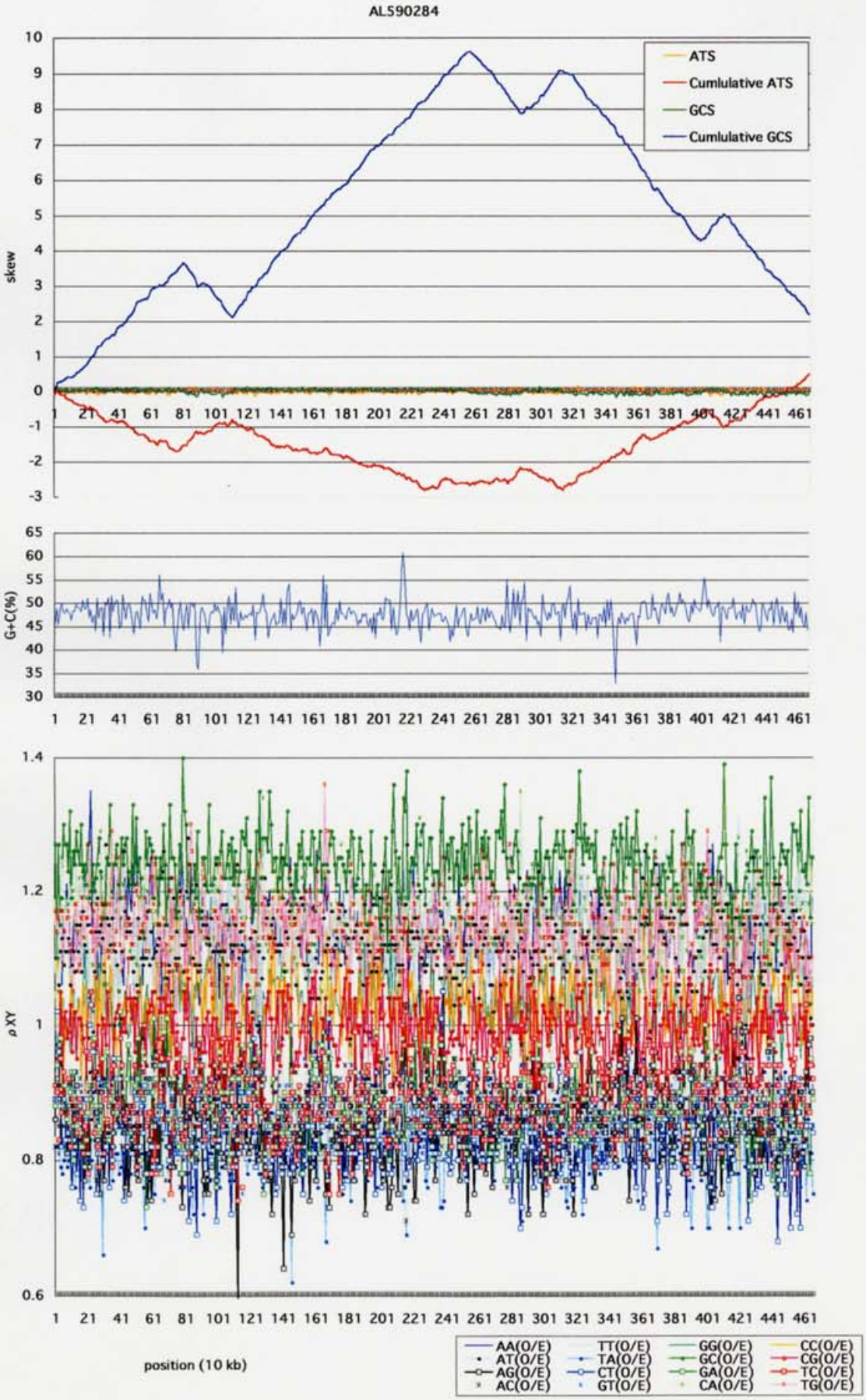
B-40 *Salmonella enterica subsp. entericaserovar Typhi* NC_003198 4,809,037 11 52.1



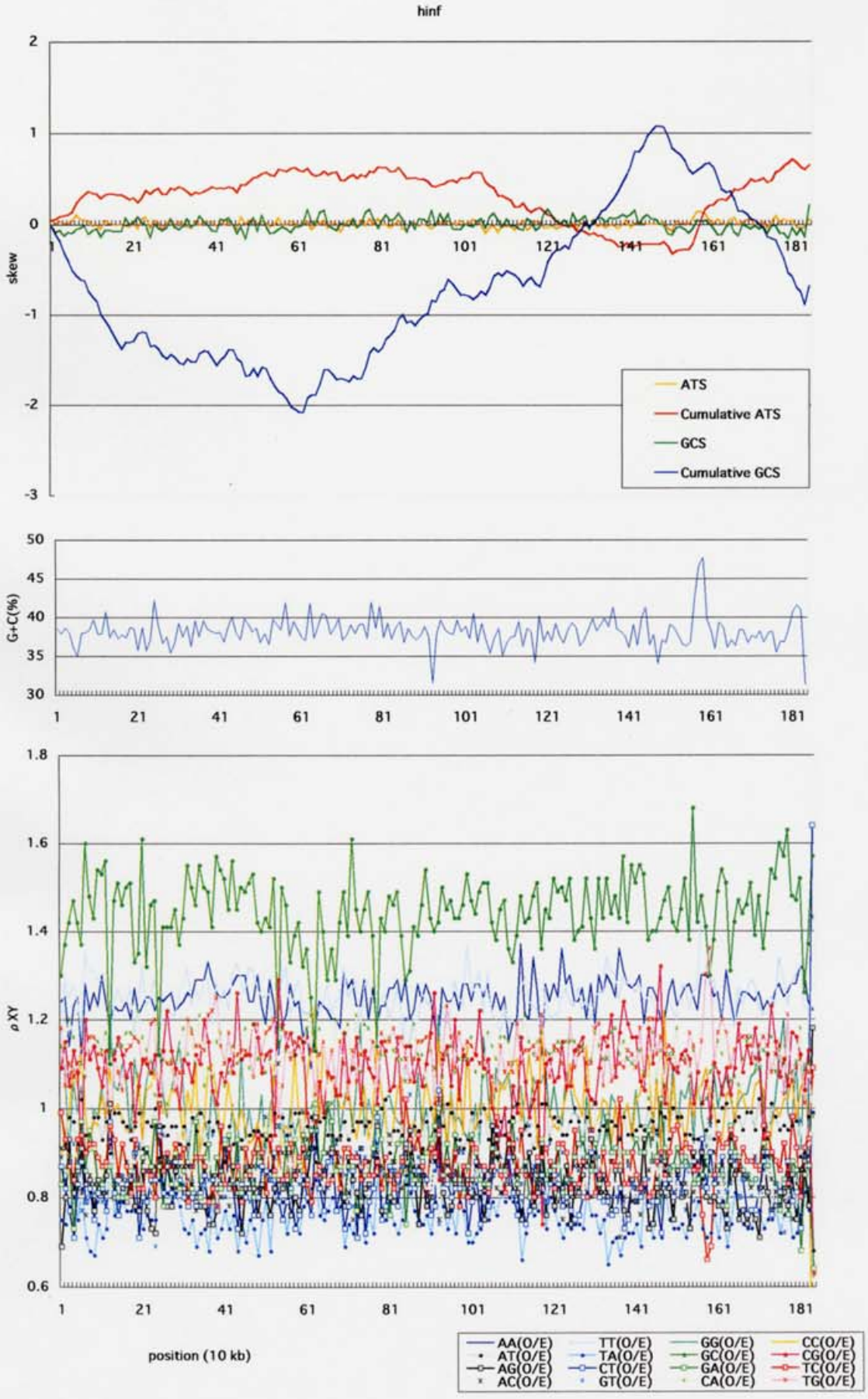
B-41 *Salmonella typhimurium LT2* NC_003197 4,857,432 11 52.2



B-55 *Yersinia pestis* NC_003143 4,653,728 11 47.6

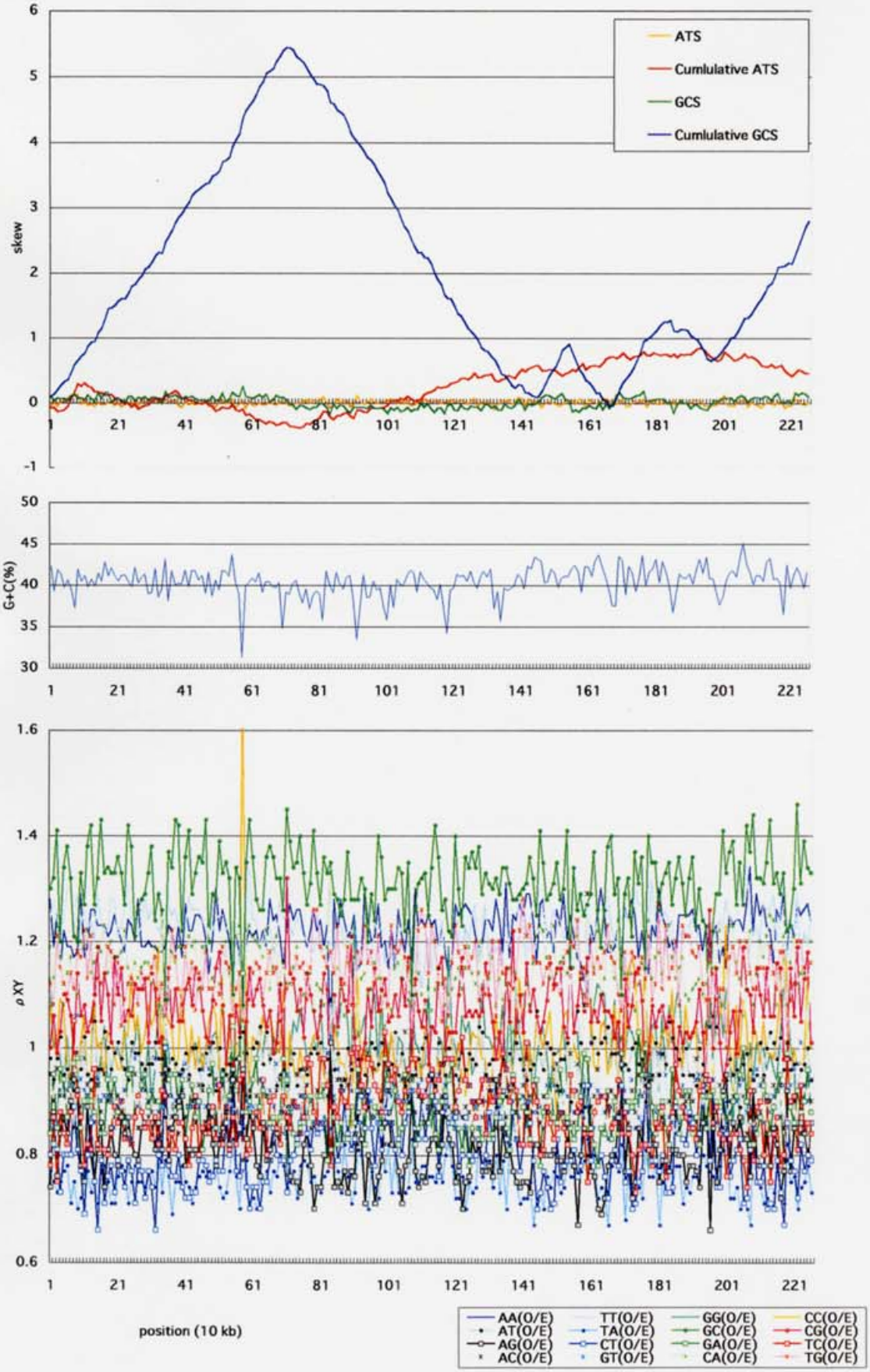


B-21 *Haemophilus influenzae Rd* NC_000907 1,830,138 11 38.2

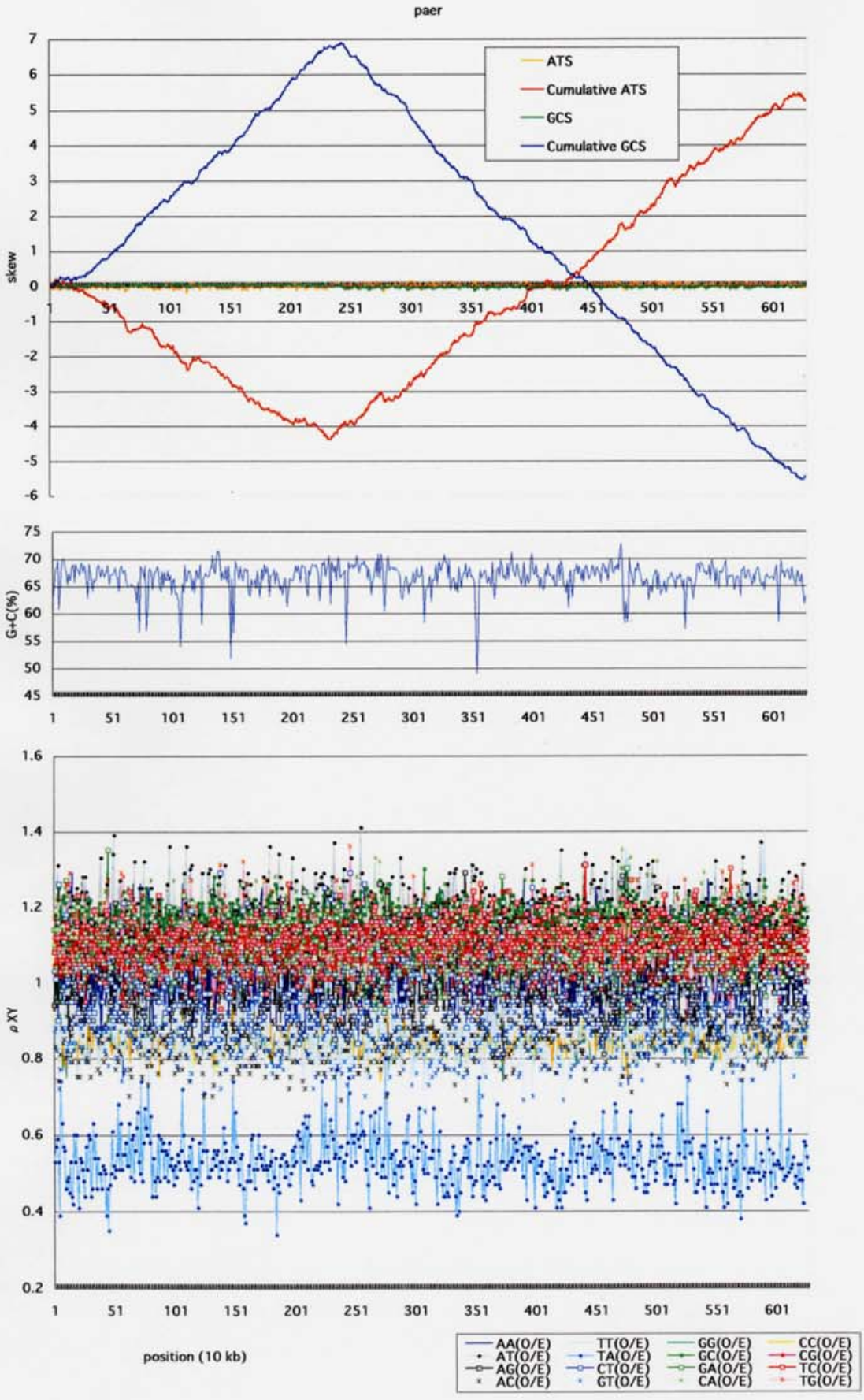


B-36 *Pasteurella multocida* NC_002663 2,257,487 11 40.4

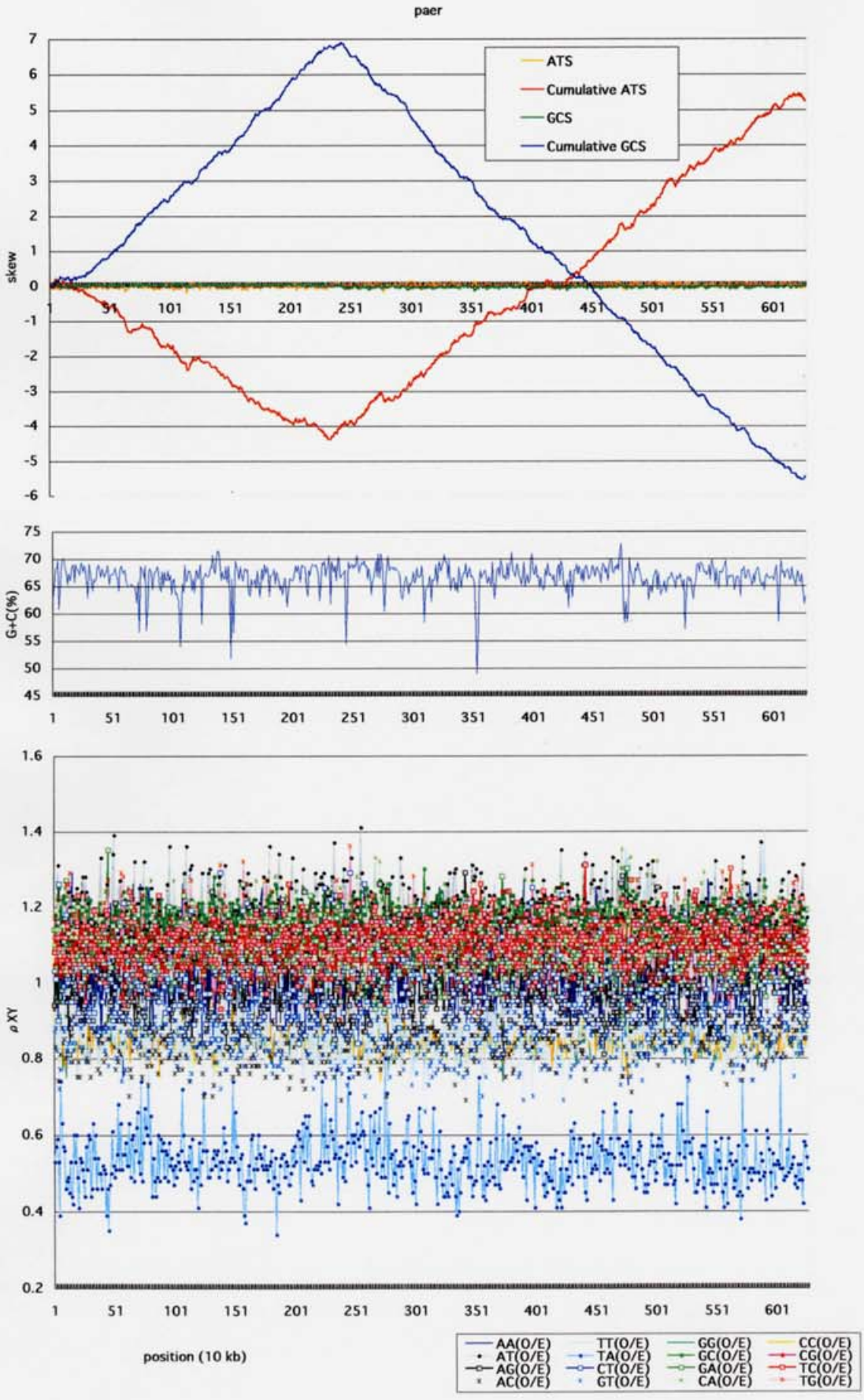
AE004439



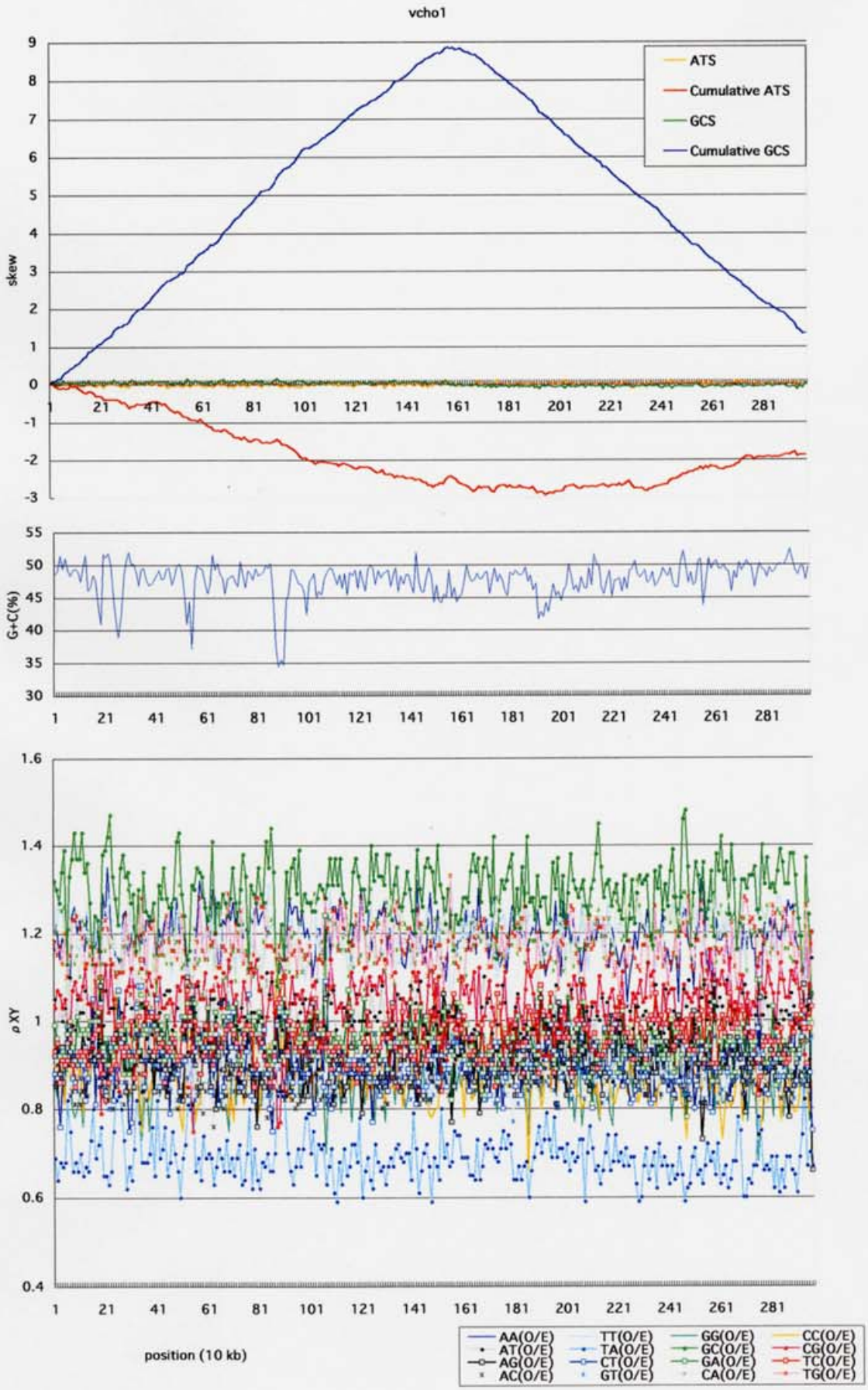
B-37 *Pseudomonas aeruginosa* NC_002516 6,264,403 11 66.6



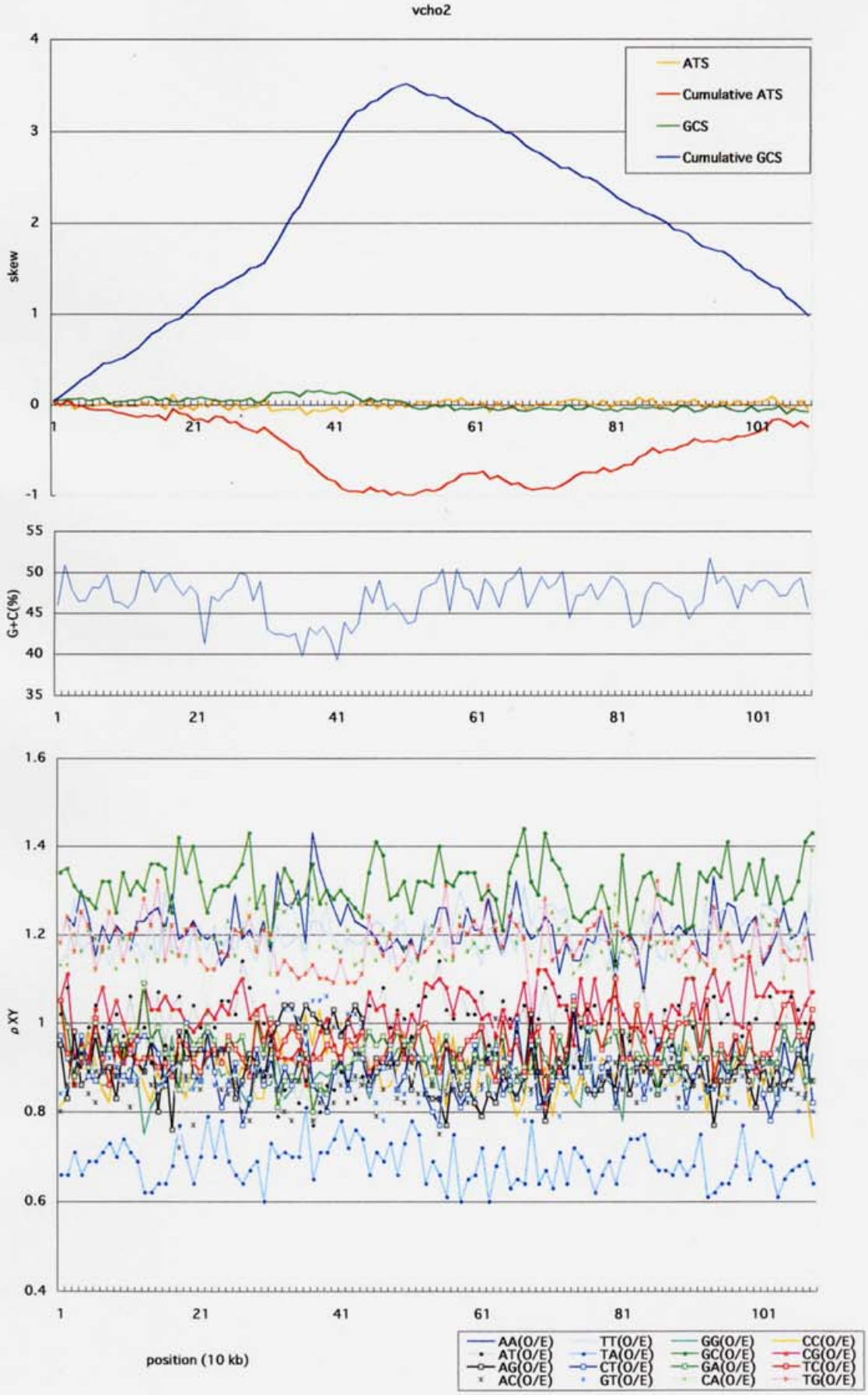
B-37 *Pseudomonas aeruginosa* NC_002516 6,264,403 11 66.6



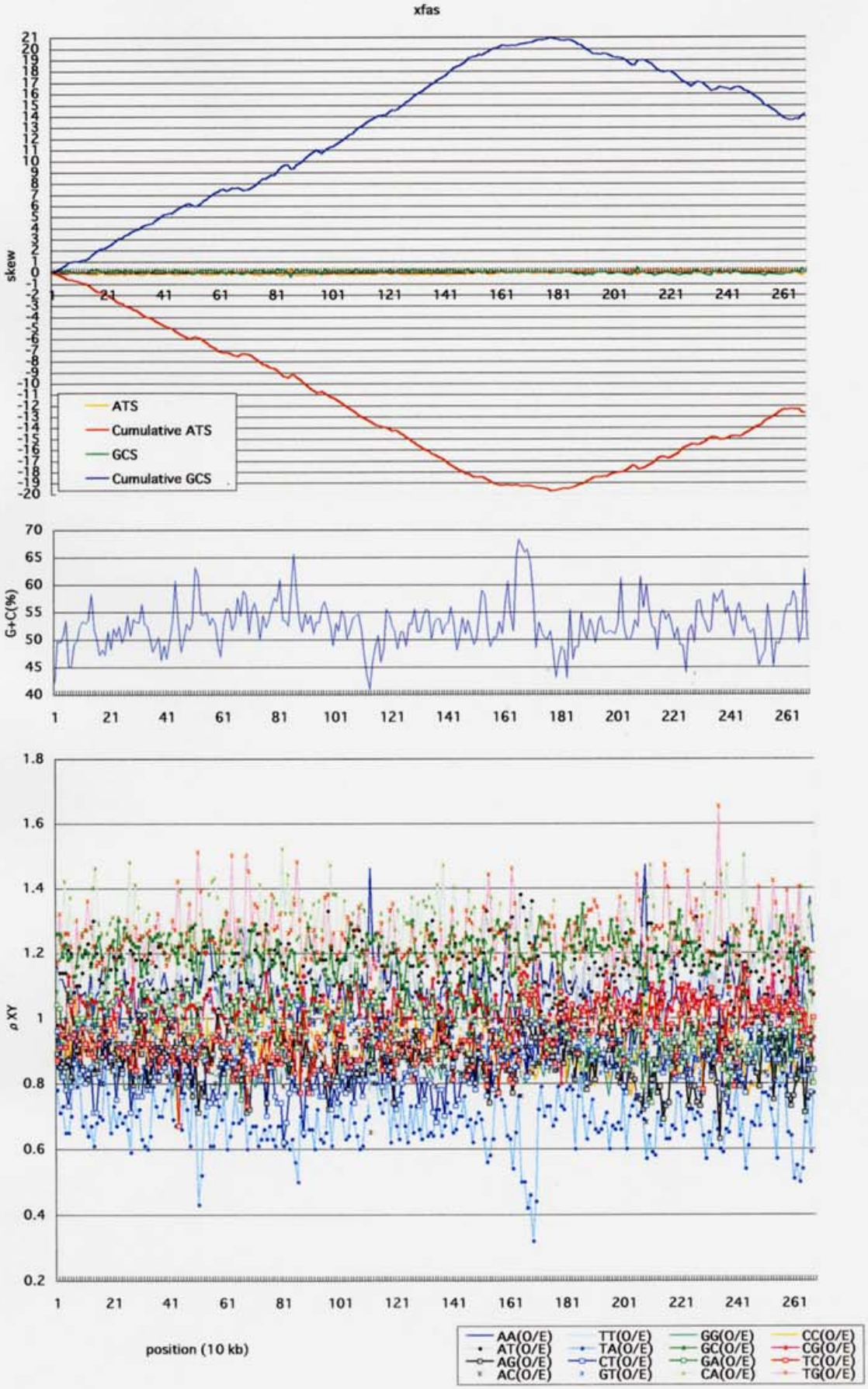
B-52 *Vibrio cholerae* chromosome1 NC_002505 2,961,149 11 47.7



B-53 *Vibrio cholerae* chromosome2 NC_002506 1,072,315 11 46.9



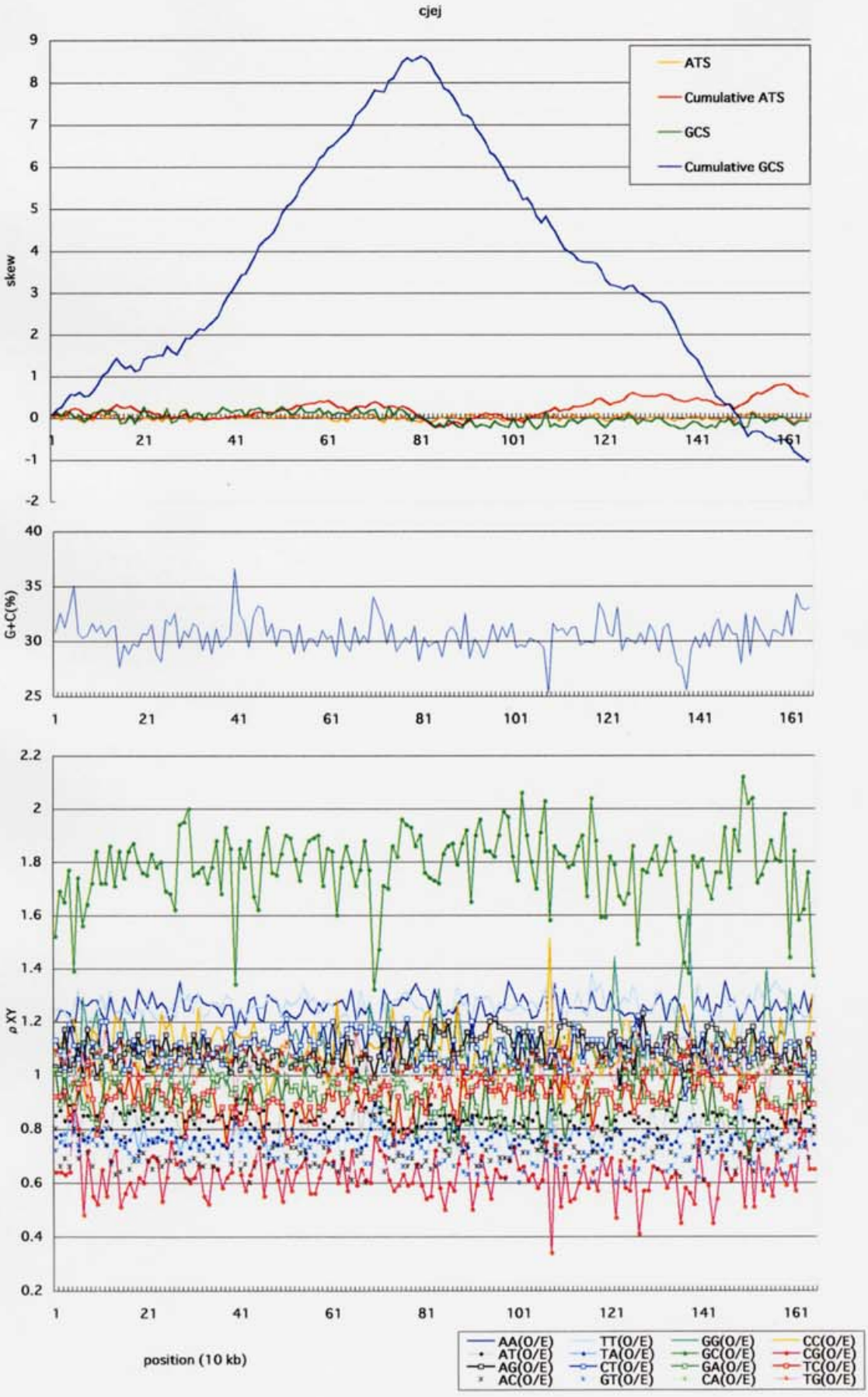
B-54 *Xylella fastidiosa 9a5c* NC_002488 2,679,306 11 52.7



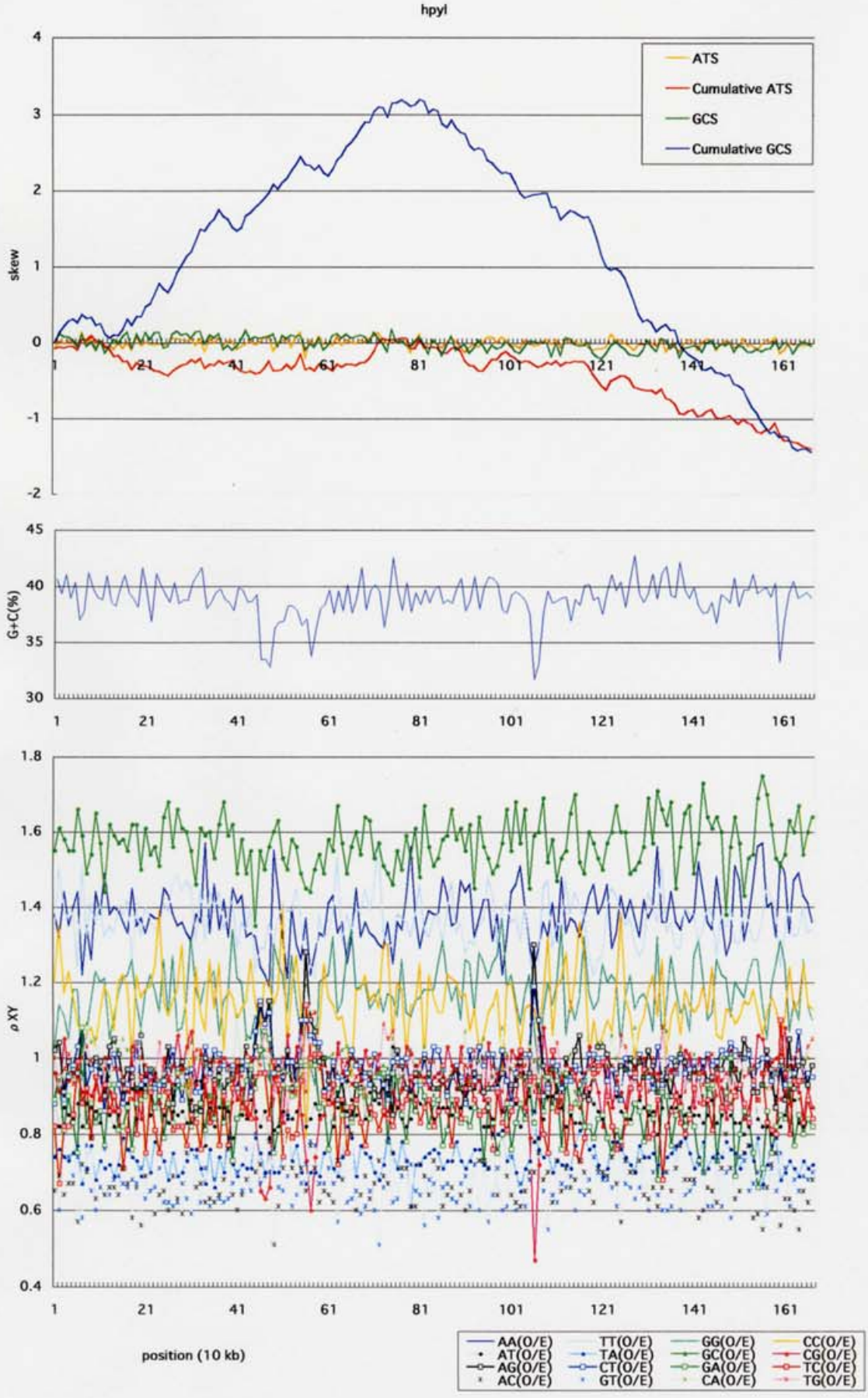
Bacteria; Proteobacteria;

epsilon subdivision (3)

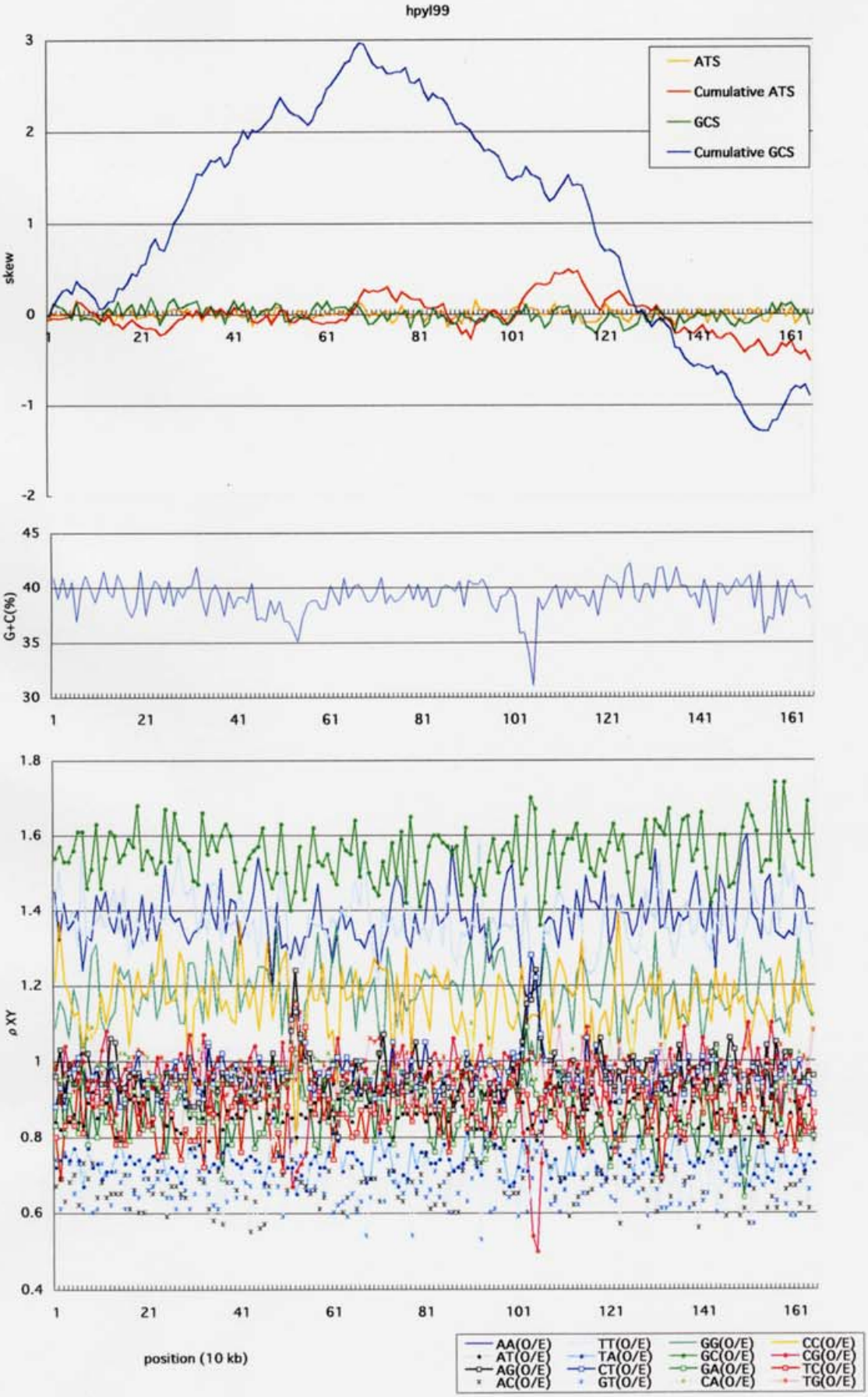
Campylobacter jejuni NC_002163 1,641,481 11 30.6



B-22 *Helicobacter pylori 26695* NC_000915 1,667,867 11 38.9



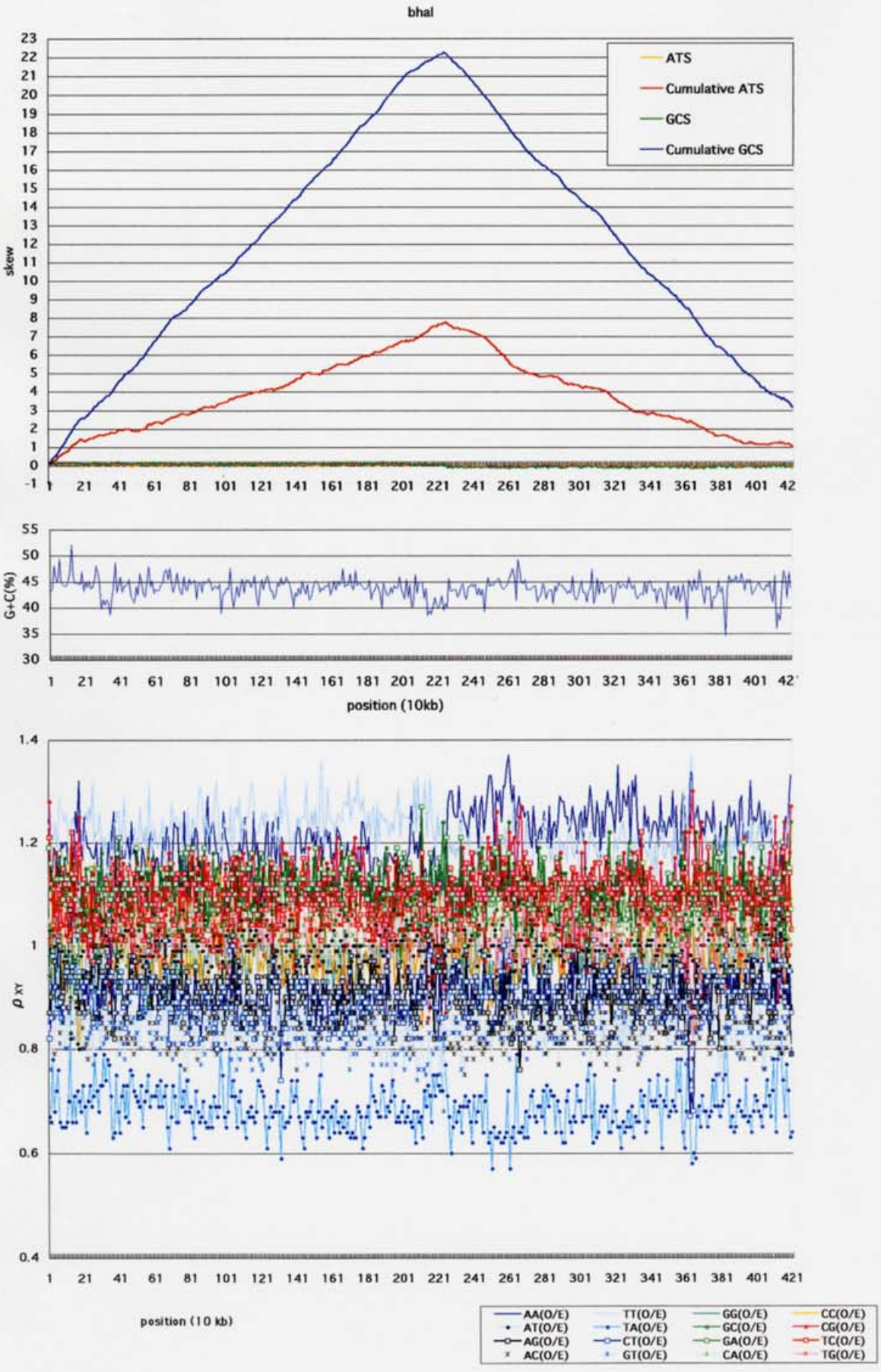
B-23 *Helicobacter pylori J99* NC_000921 1,643,831 11 39.2



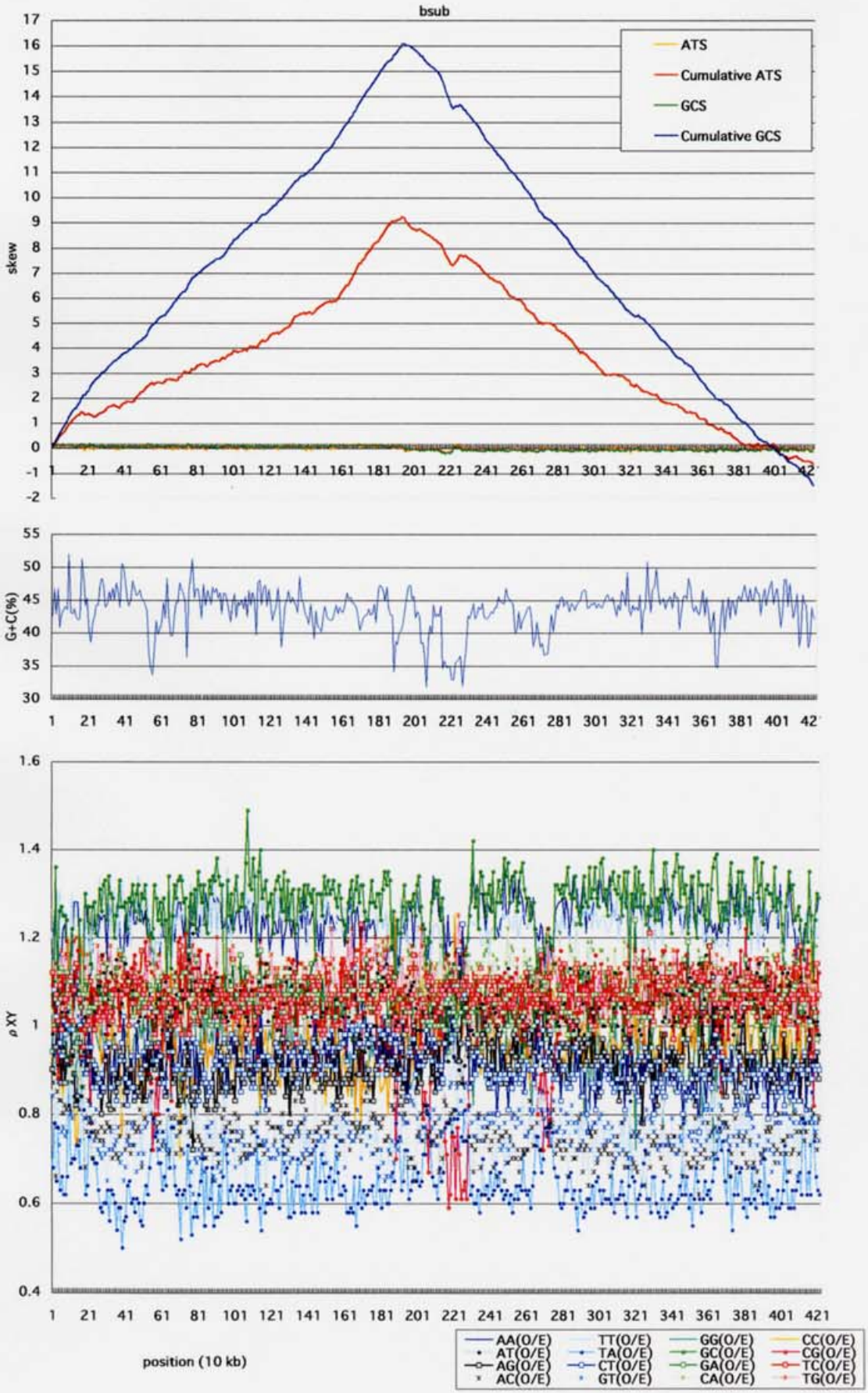
Bacteria; Firmicutes;

Bacillus/Clostridium group (15)

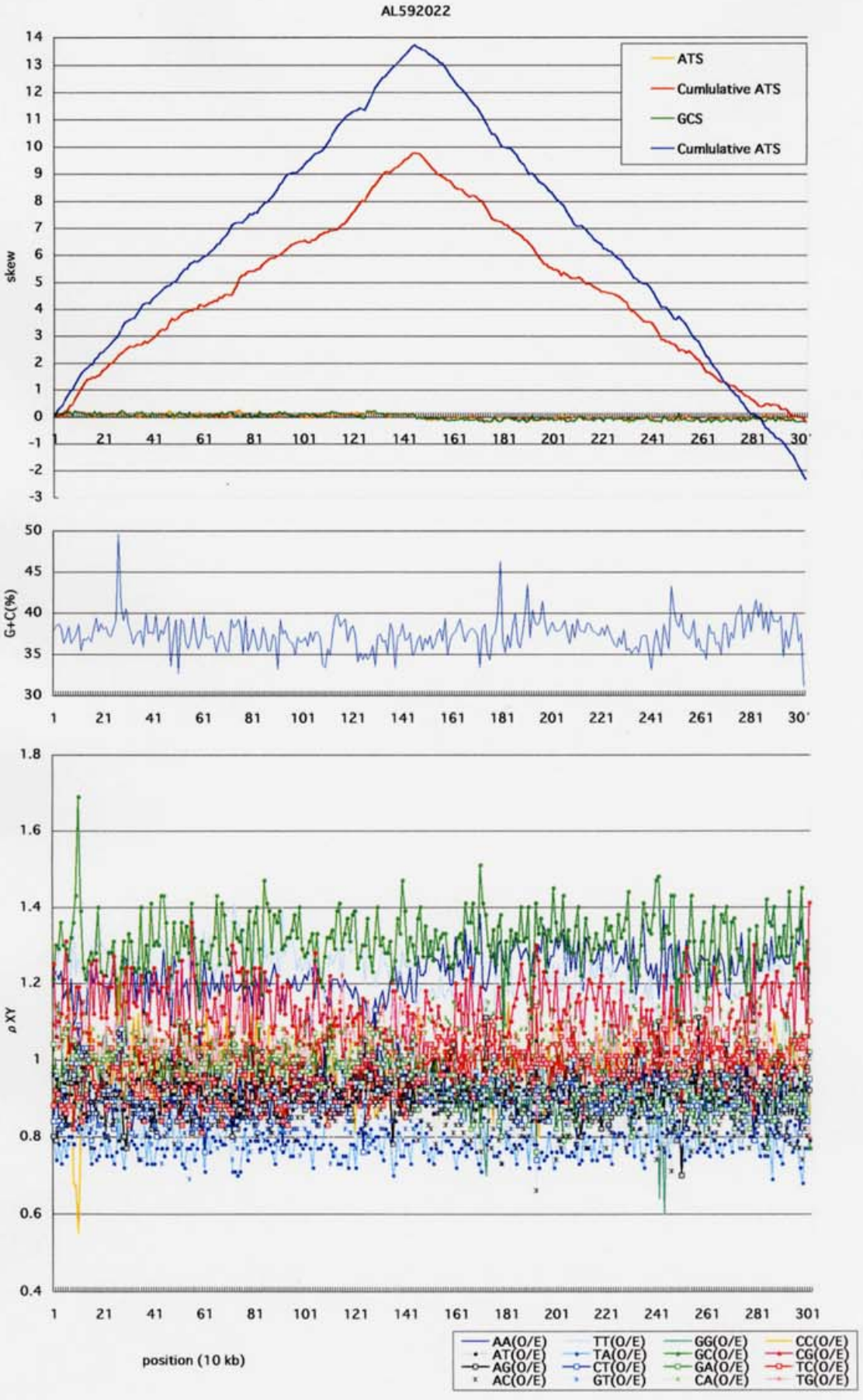
B—4 *Bacillus halodurans* NC_002570 4,202,353 11 43.7



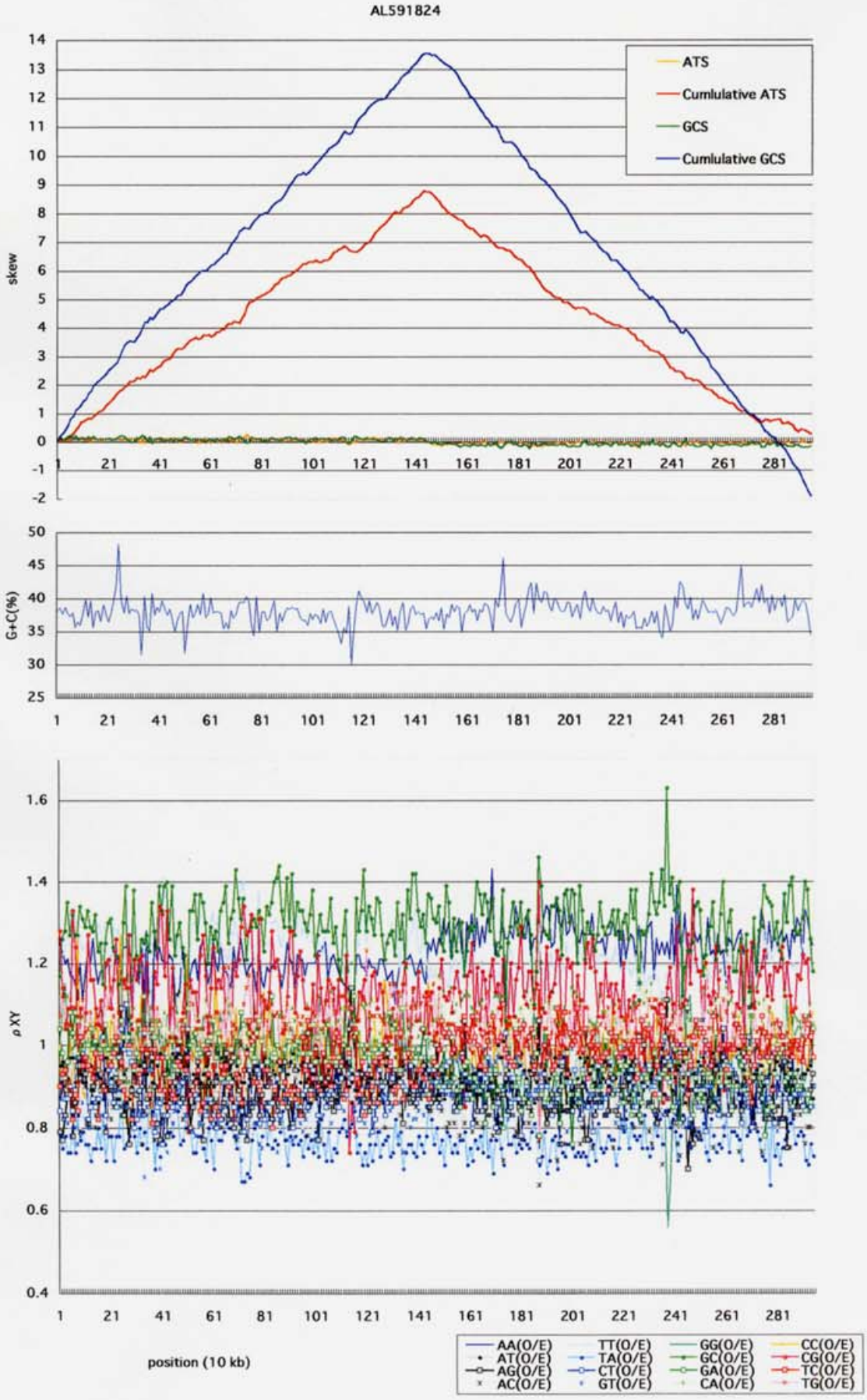
B—5 *Bacillus subtilis* NC_000964 4,214,814 11 43.5



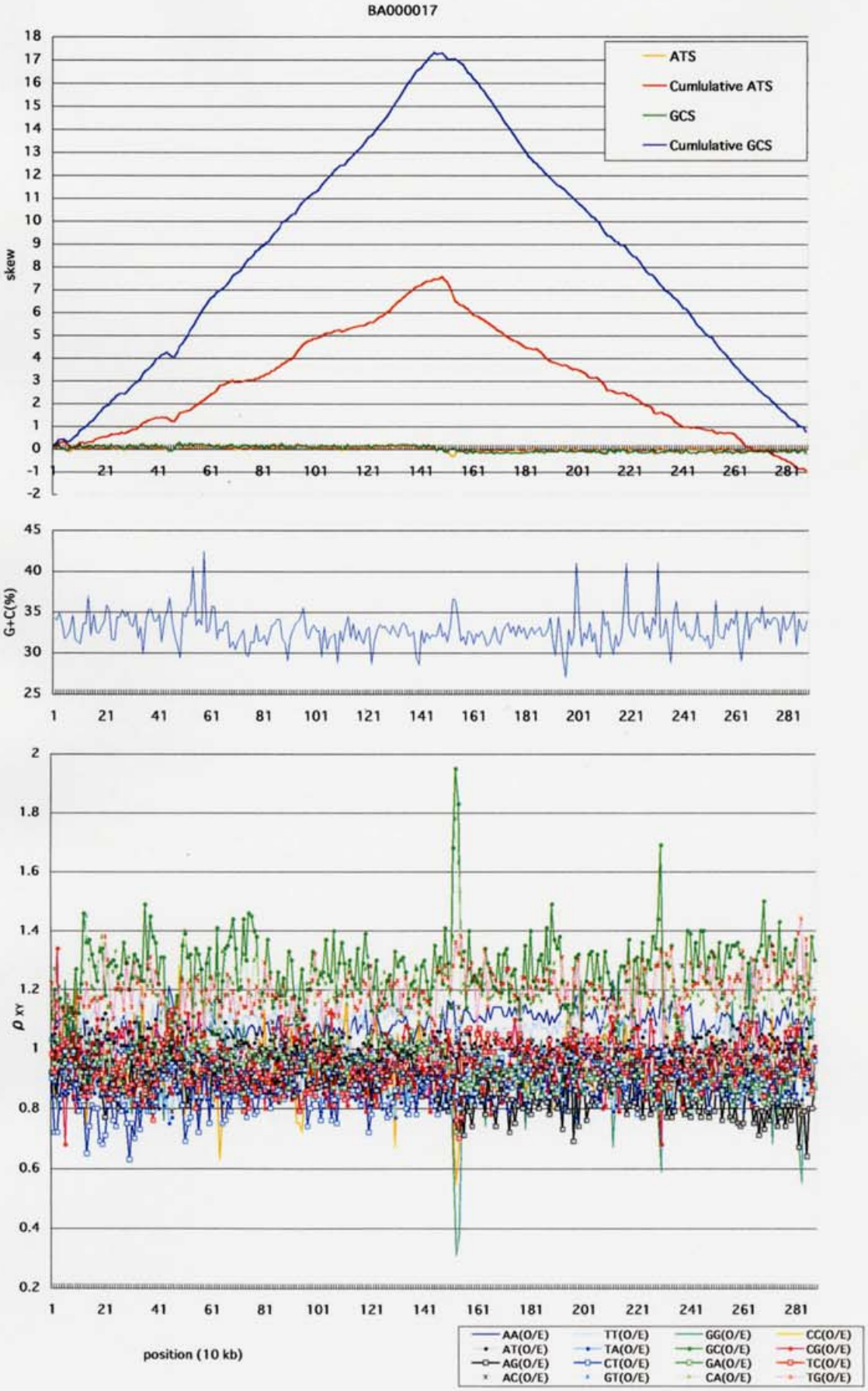
B-25 *Listeria innocua* NC_003212 3,011,208 11 37.4



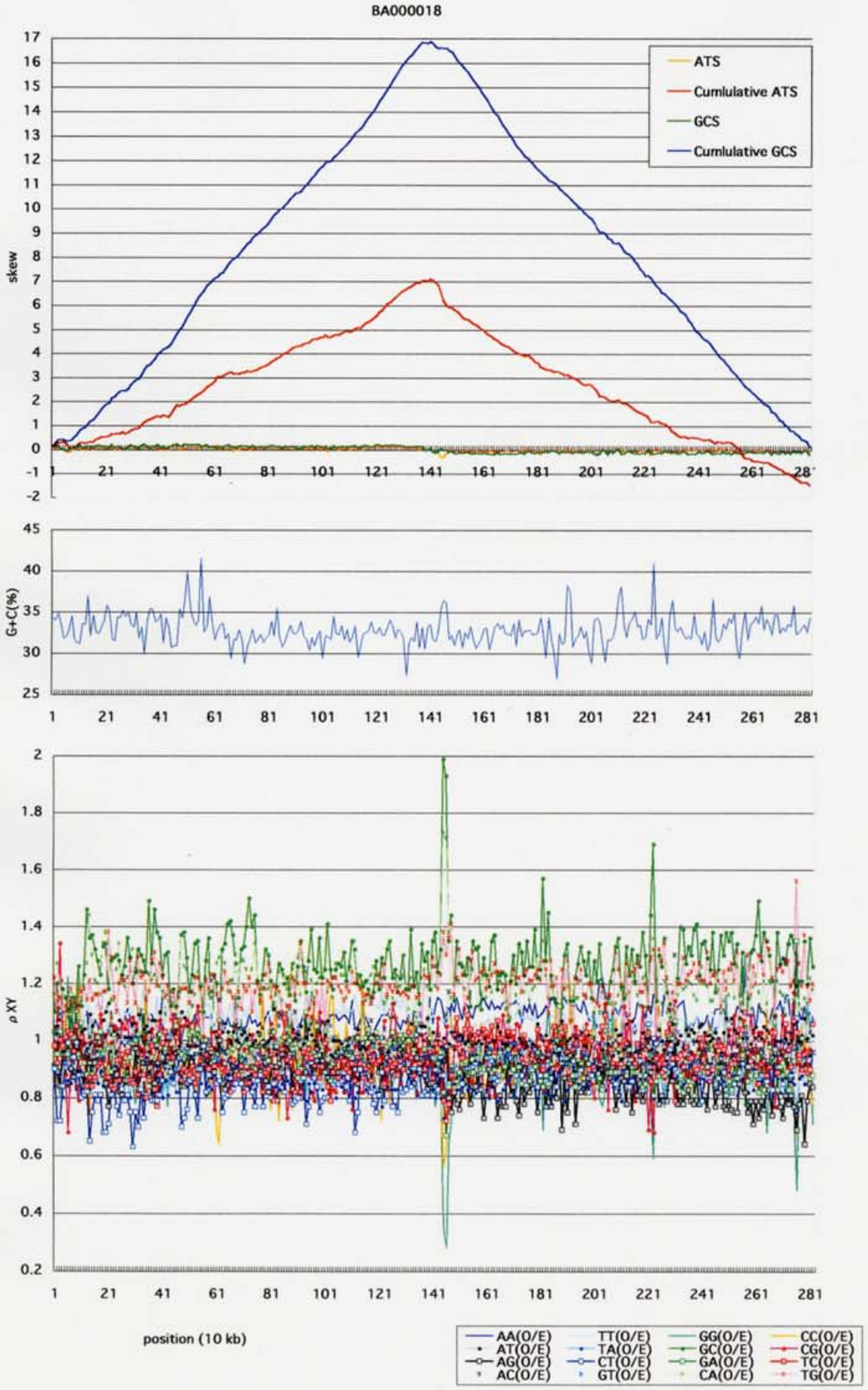
B-26 *Listeria monocytogenes EGD-e* NC_003210 2,944,528 11 38.0



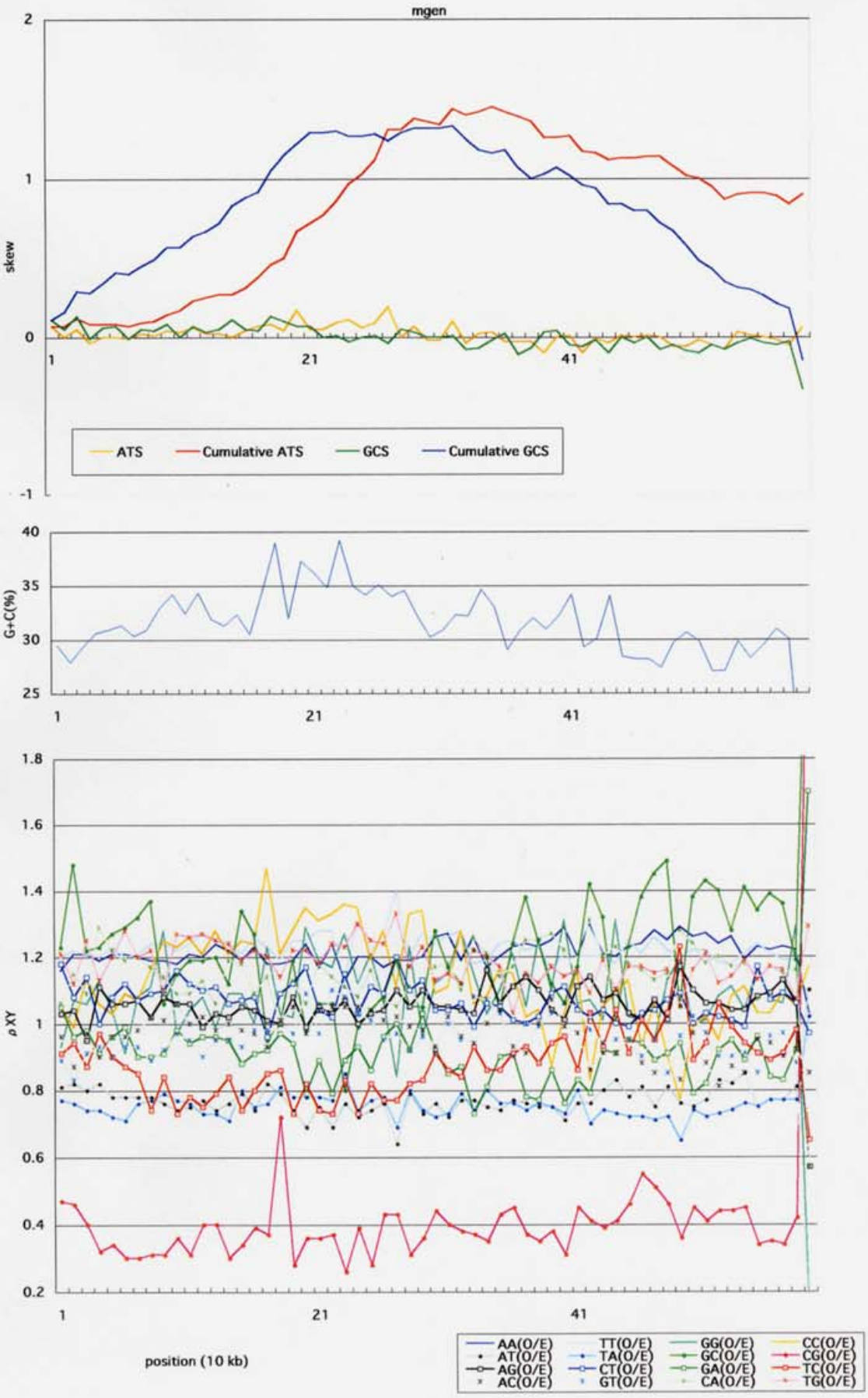
B-43 *Staphylococcus aureus subsp. AureusMu50* NC_002758 2,878,040 11 32.9



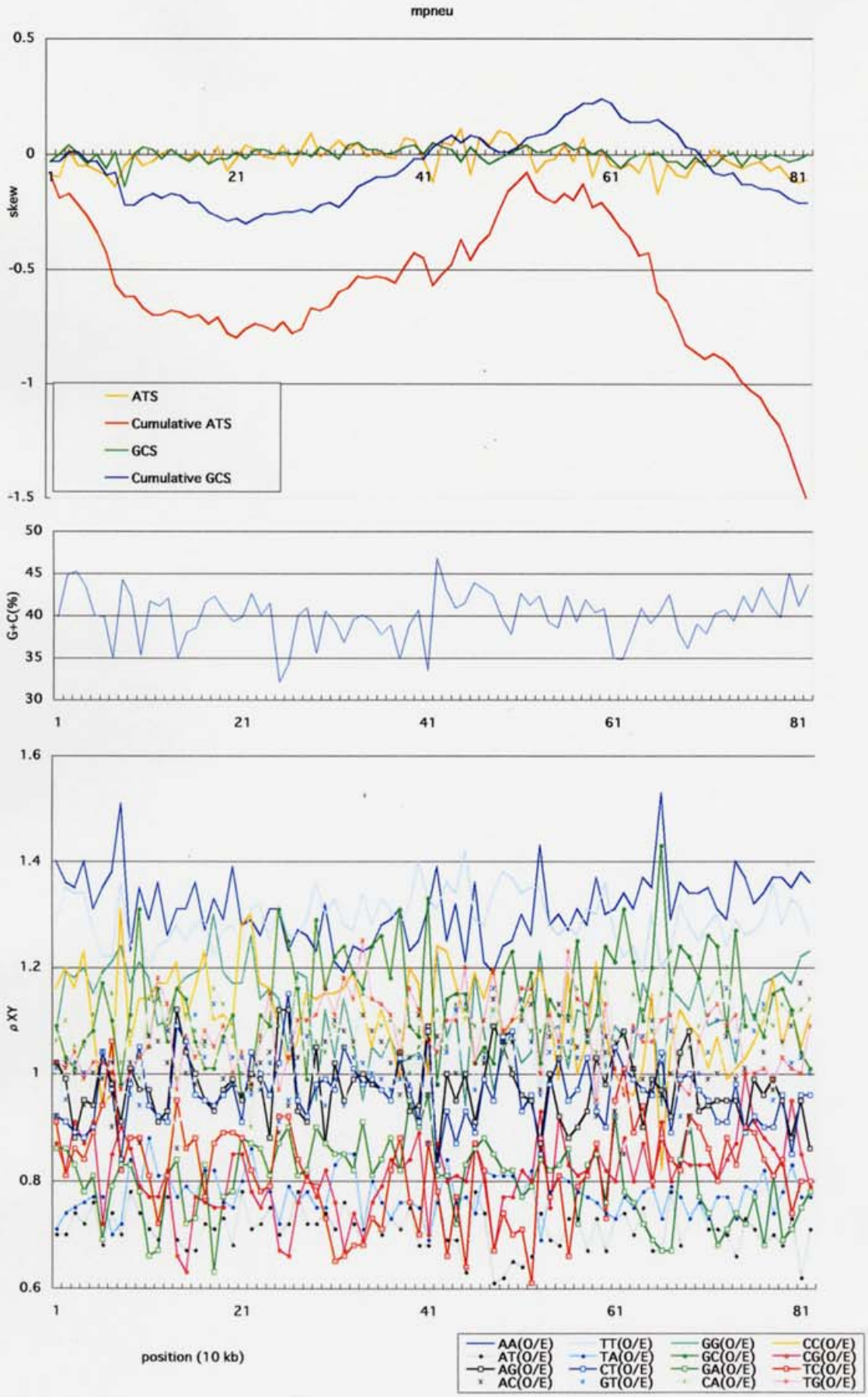
B-44 *Staphylococcus aureus subsp. aureusN315* NC_002745 2,813,641 11 32.8



B-31 *Mycoplasma genitalium* NC_000908 580,074 4 31.7

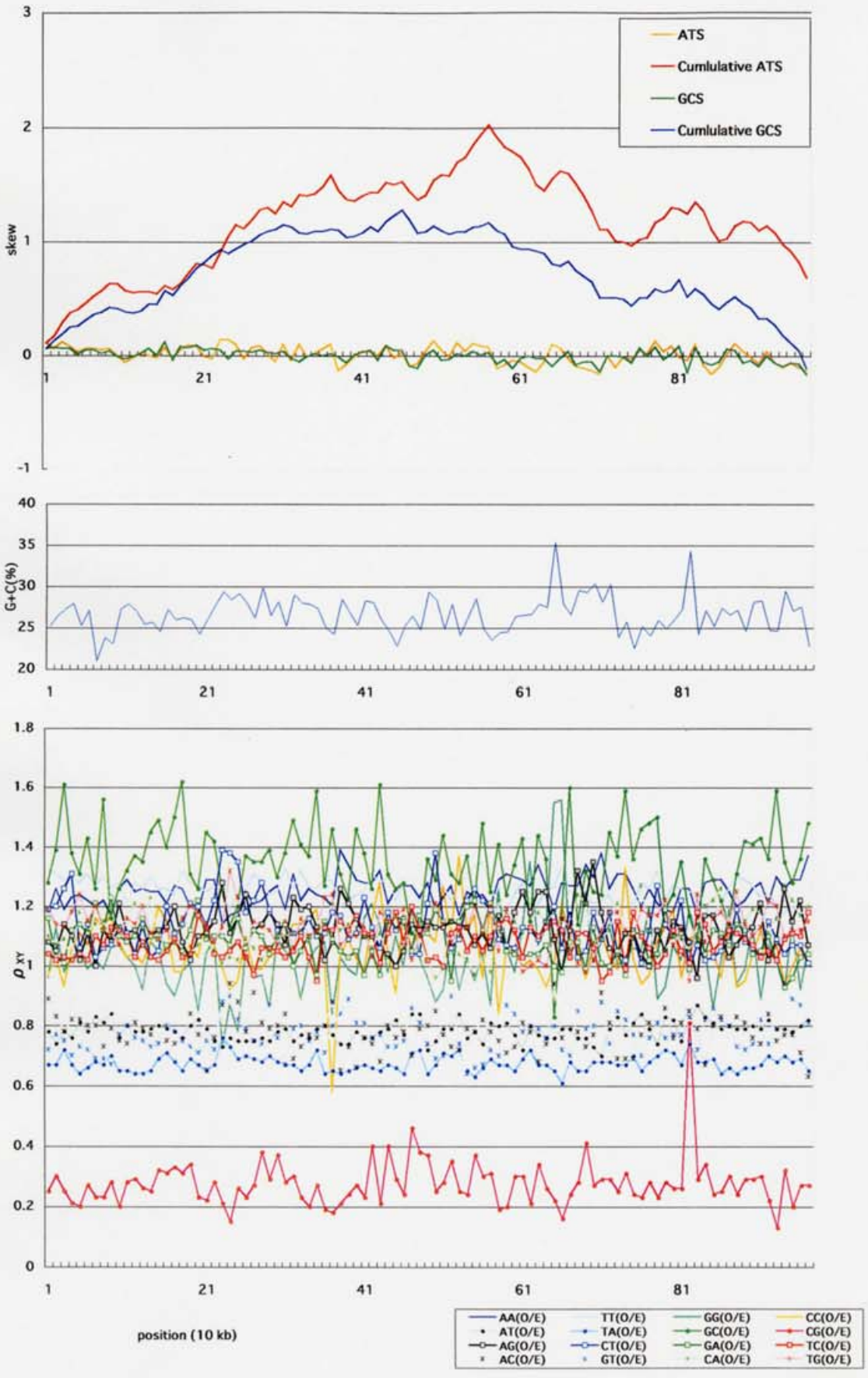


B-32 *Mycoplasma pneumoniae* NC_000912 816,394 4 40.0

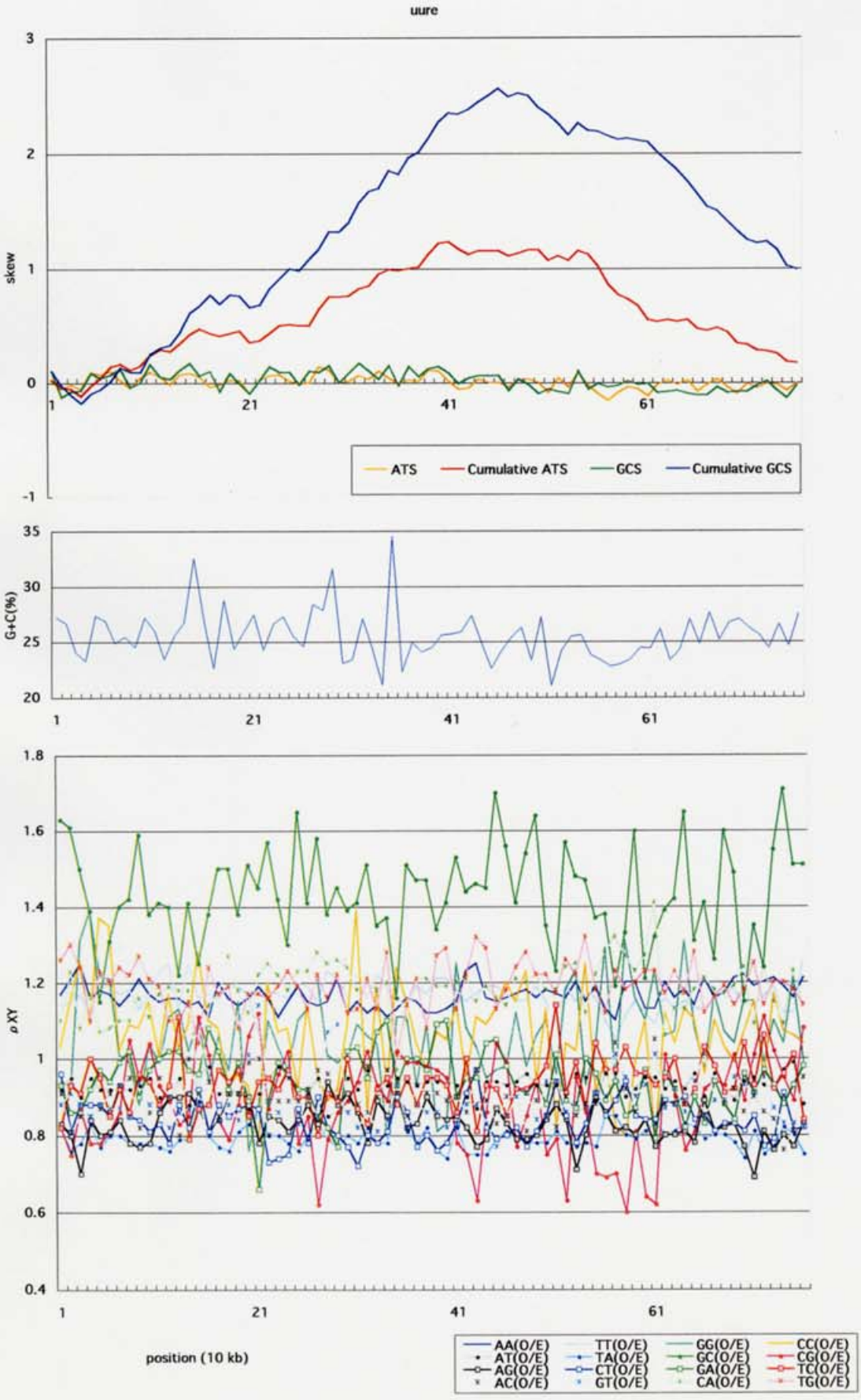


B-33 *Mycoplasma pulmonis* NC_002771 963,879 4 26.6

AL445566

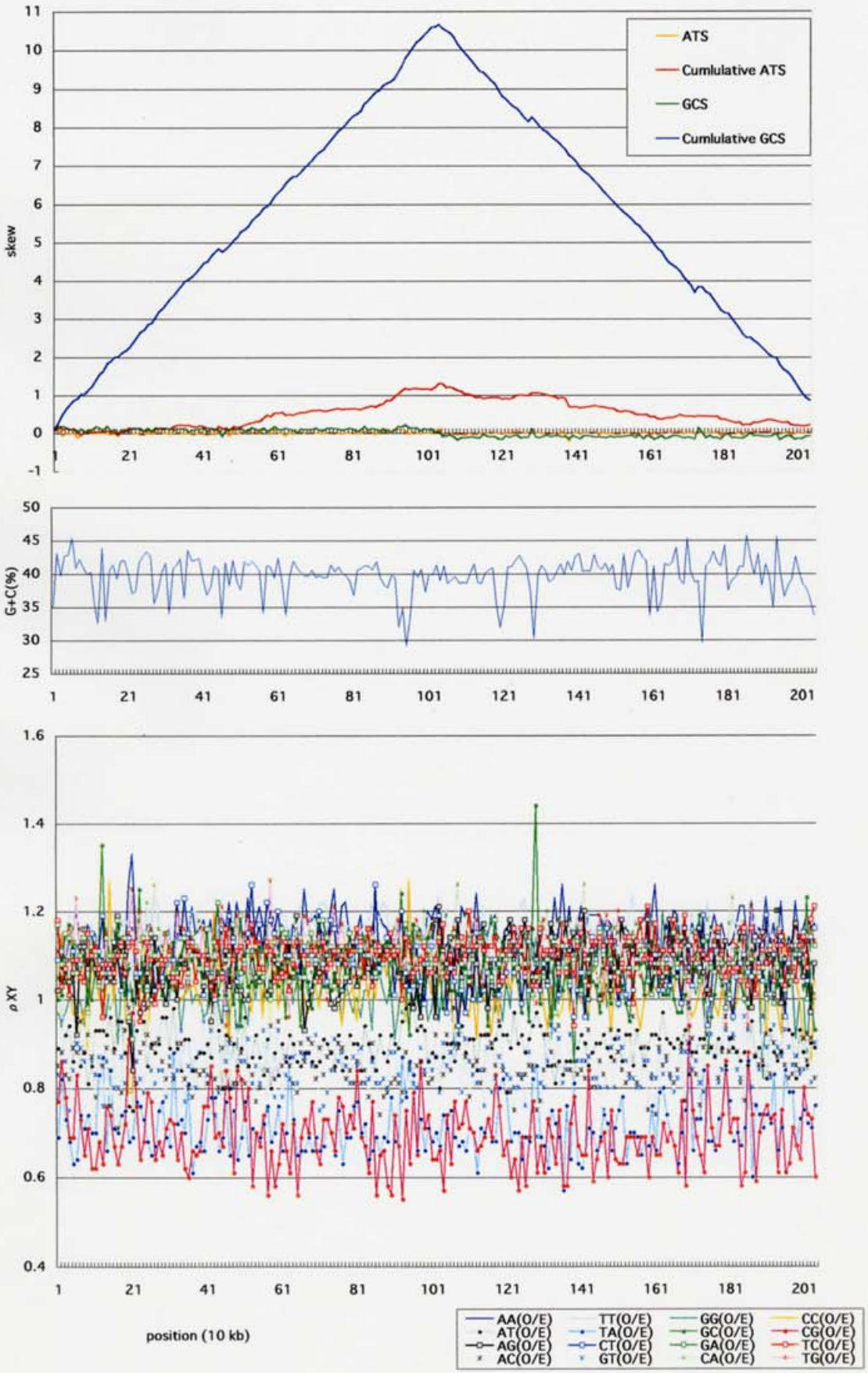


B-51 *Ureaplasma urealyticum* NC_002162 751,719 4 25.5

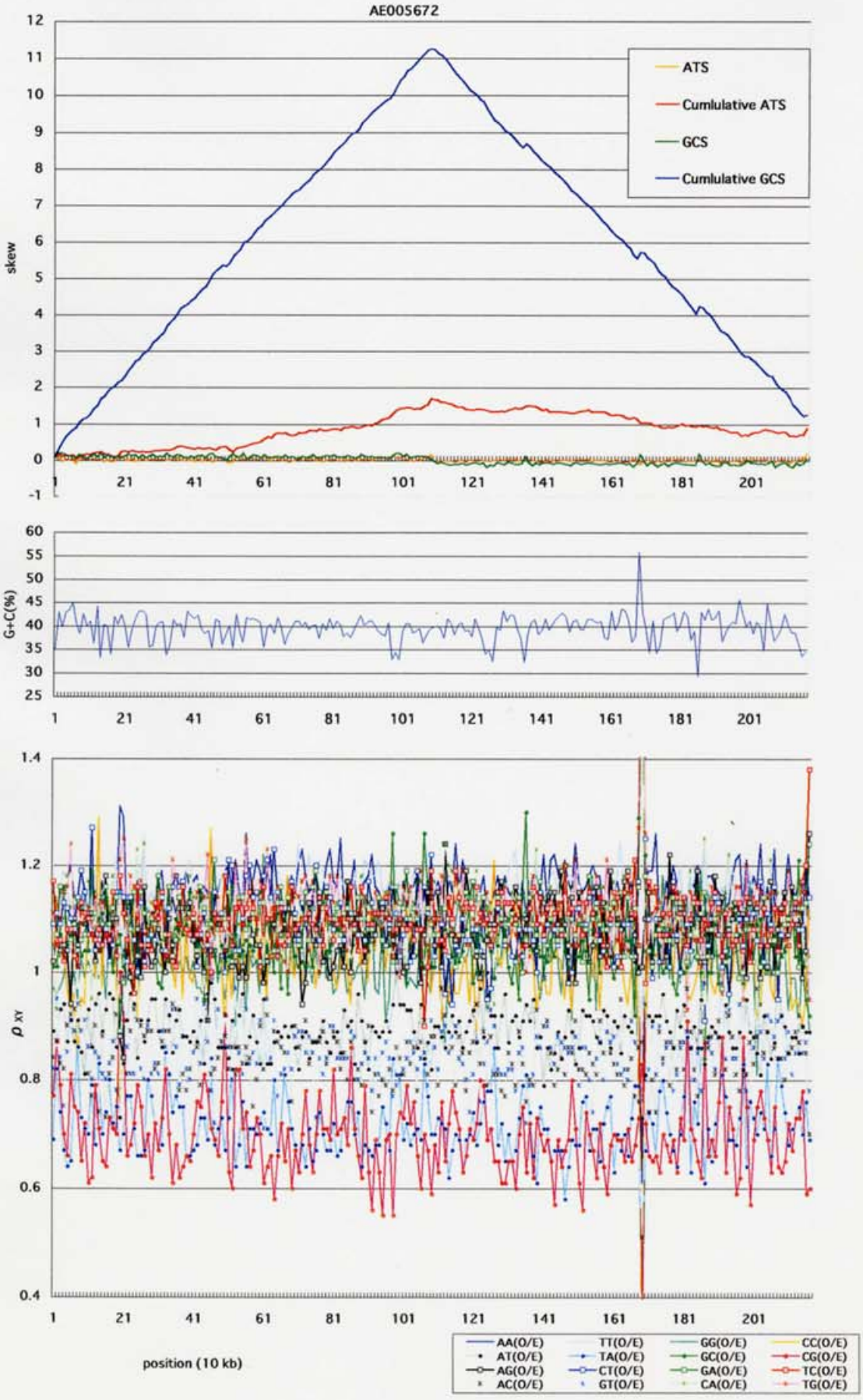


B-45 *Streptococcus pneumoniae R6* NC_003098 2,038,615 11 39.7

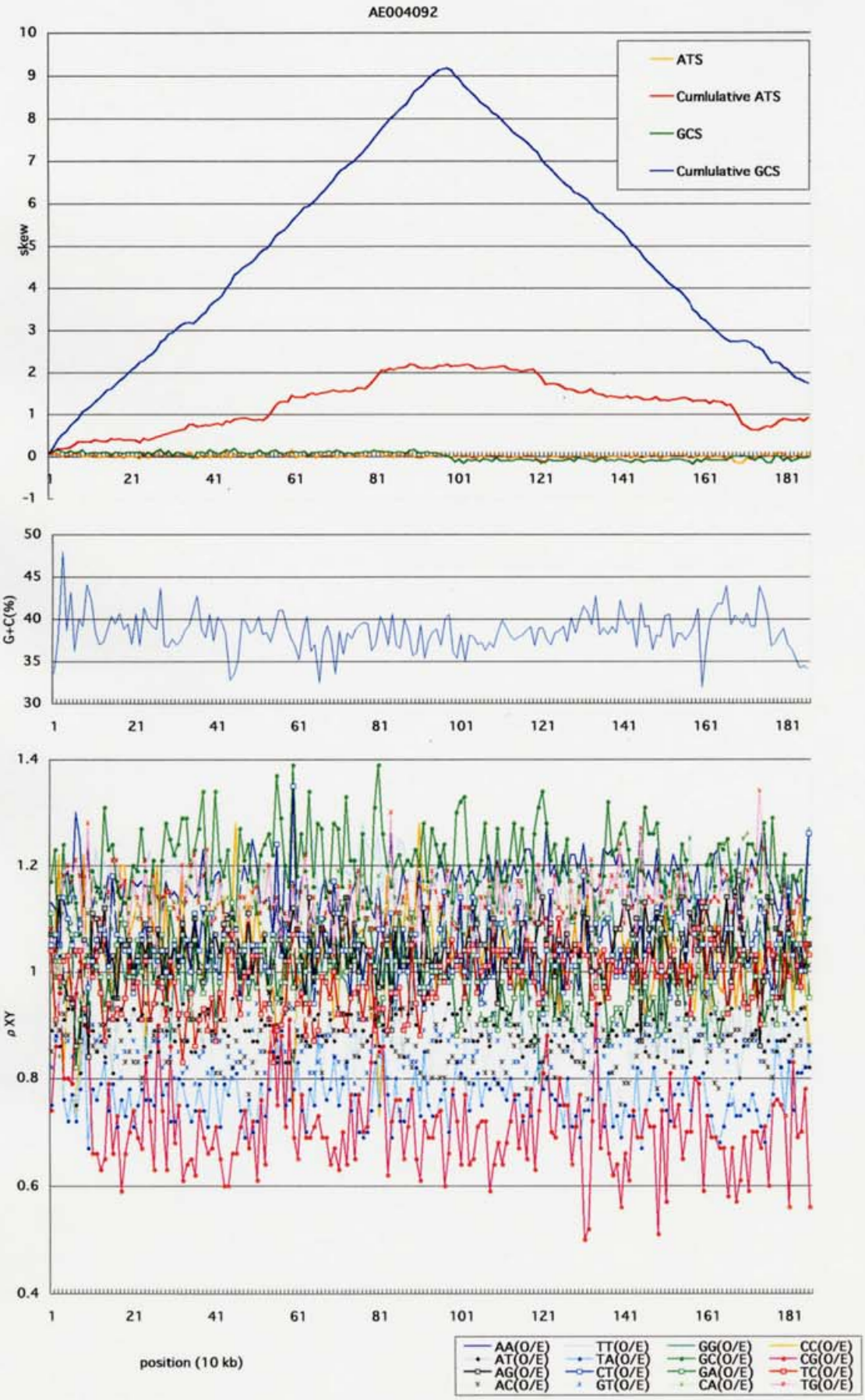
AE007317



B-46 *Streptococcus pneumoniae TIGR4* NC_003028 2,160,837 11 39.7

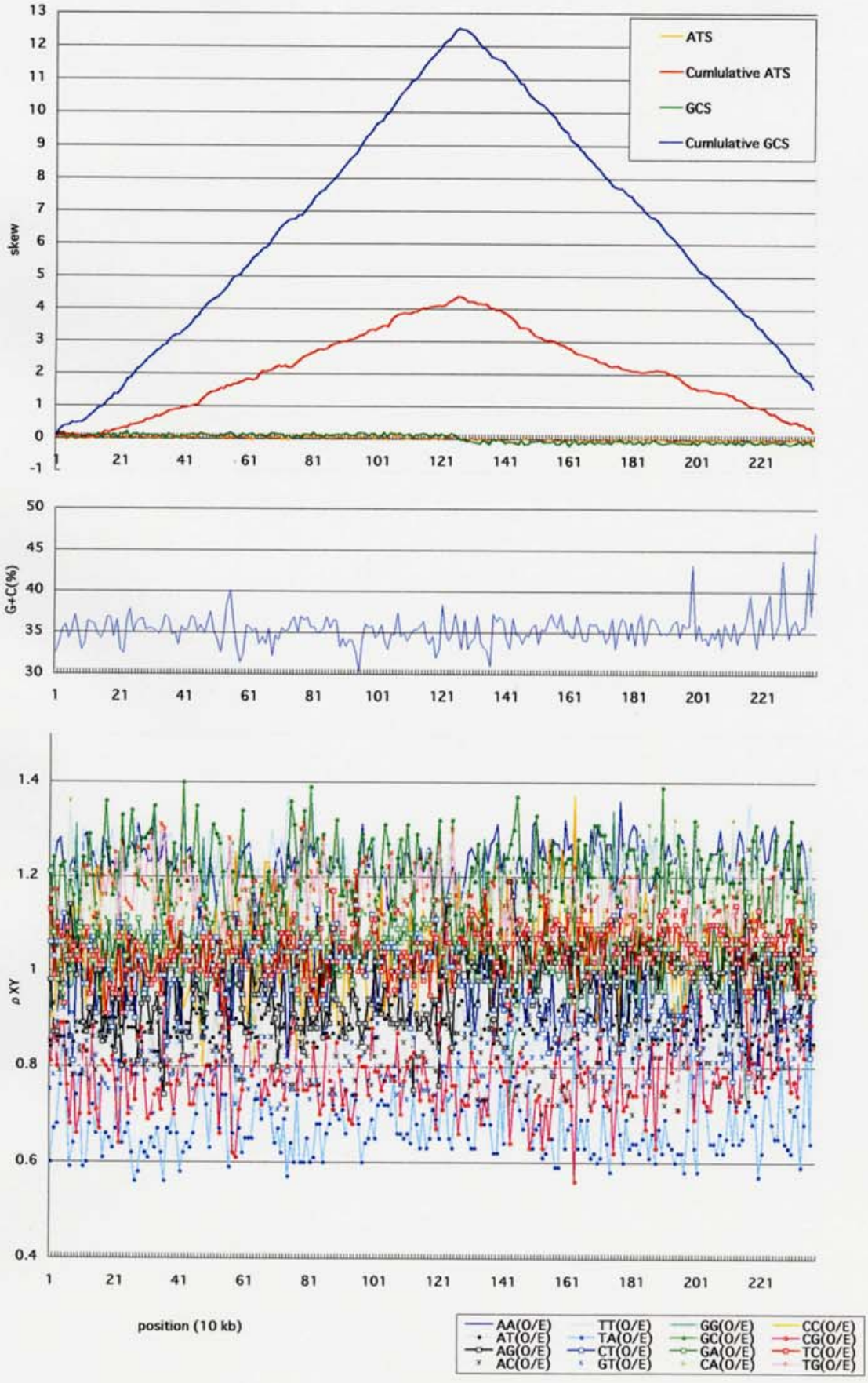


B-47 *Streptococcus pyogenes* NC_002737 1,852,441 11 38.5



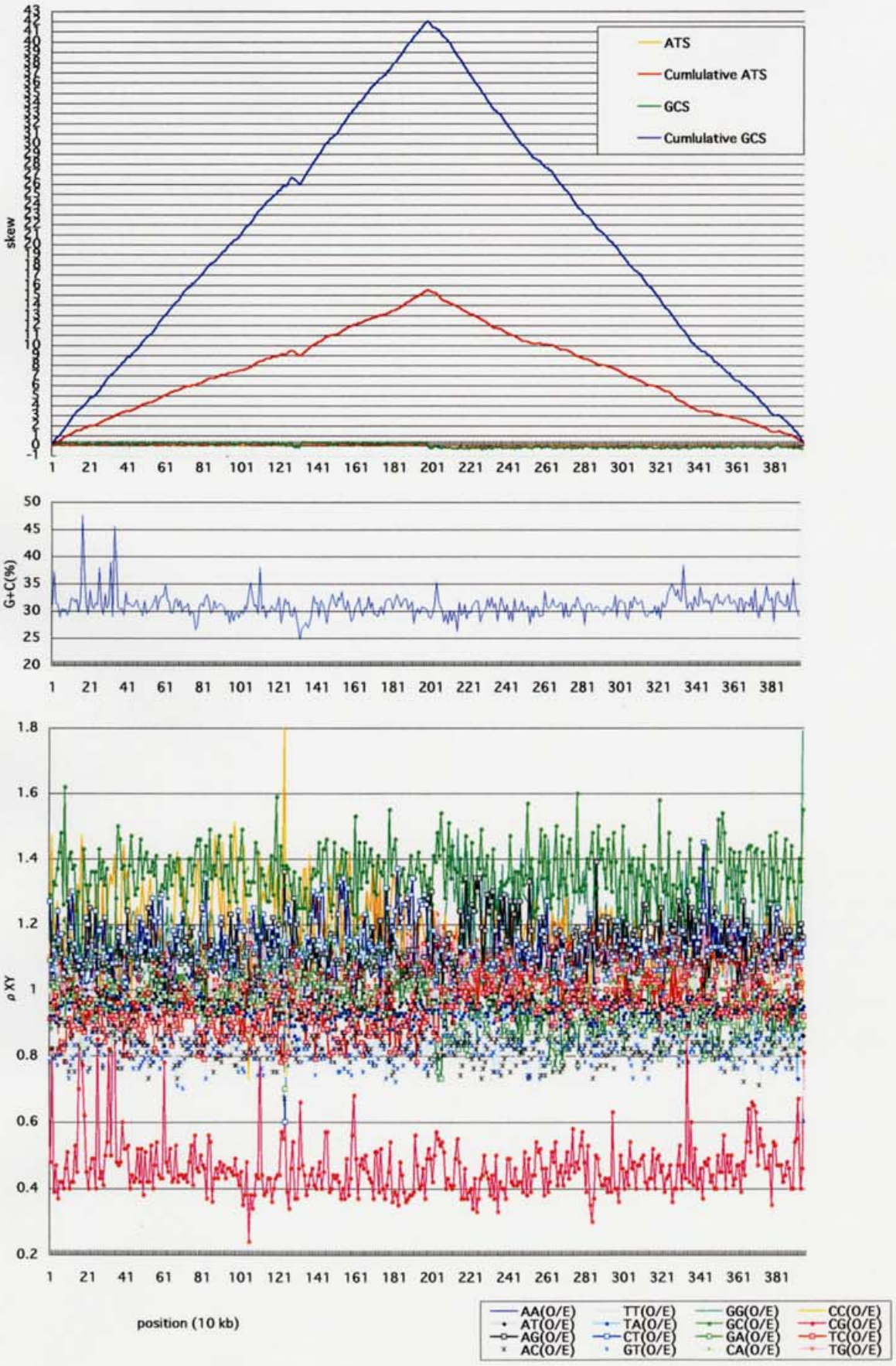
B-24 *Lactococcus lactis subsp. lactis* NC_002662 2,365,589 11 35.3

AE005176



B-15 *Clostridium acetobutylicum* NC_003030 3,940,880 11 30.9

AE001437

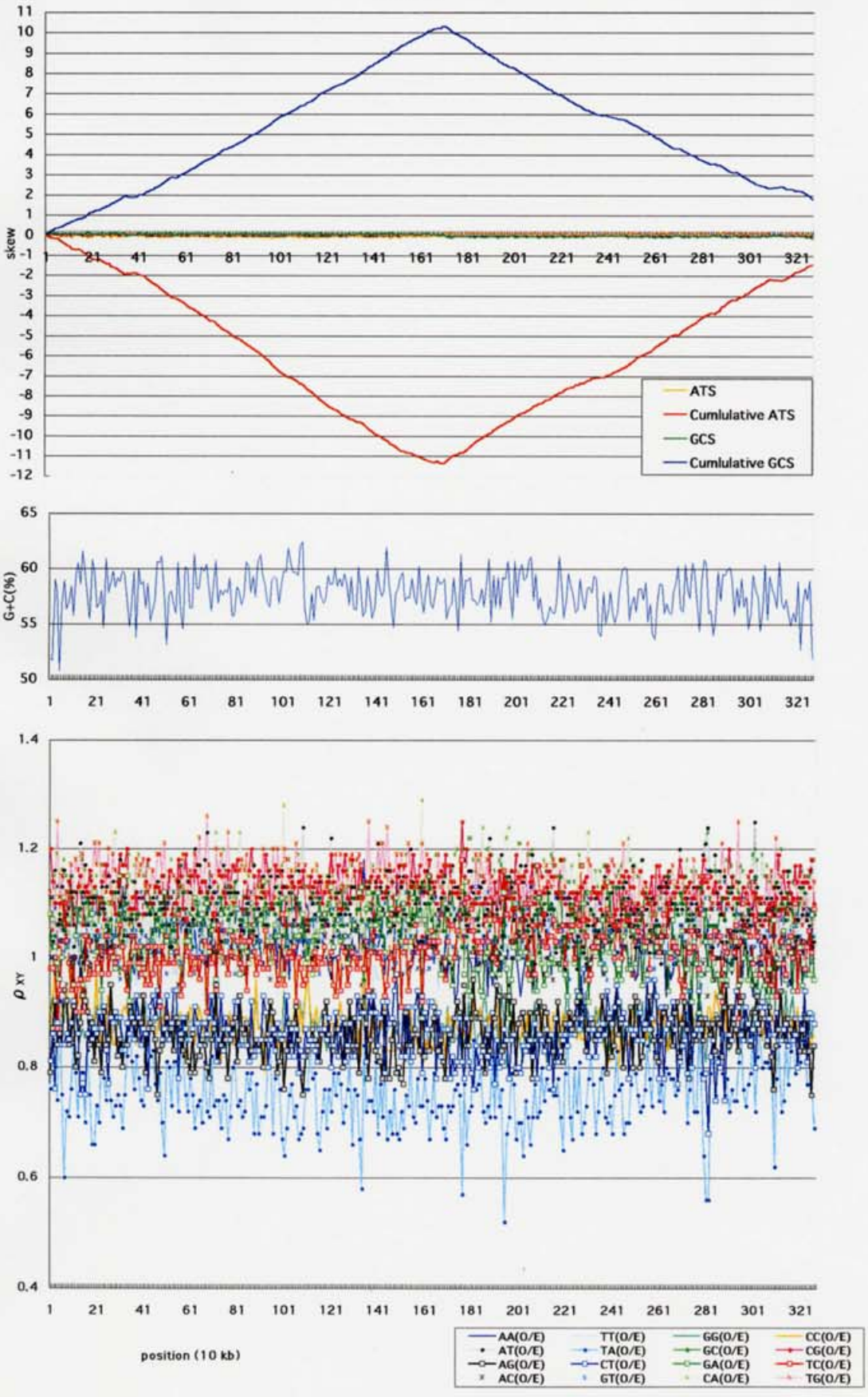


Bacteria; Firmicutes;

Actinobacteria (3)

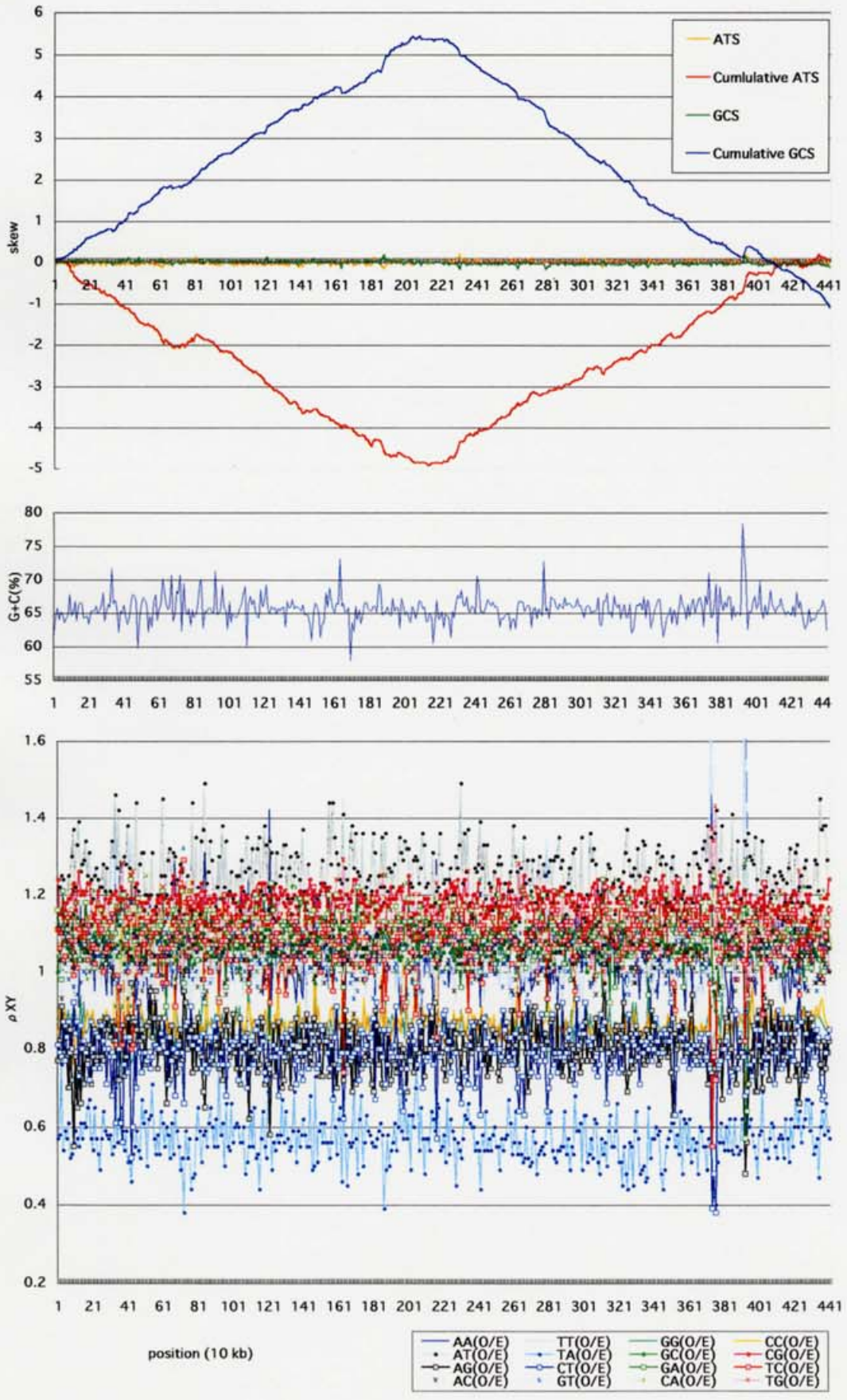
B-28 *Mycobacterium leprae* NC_002677 3,268,203 11 57.8

AL450380

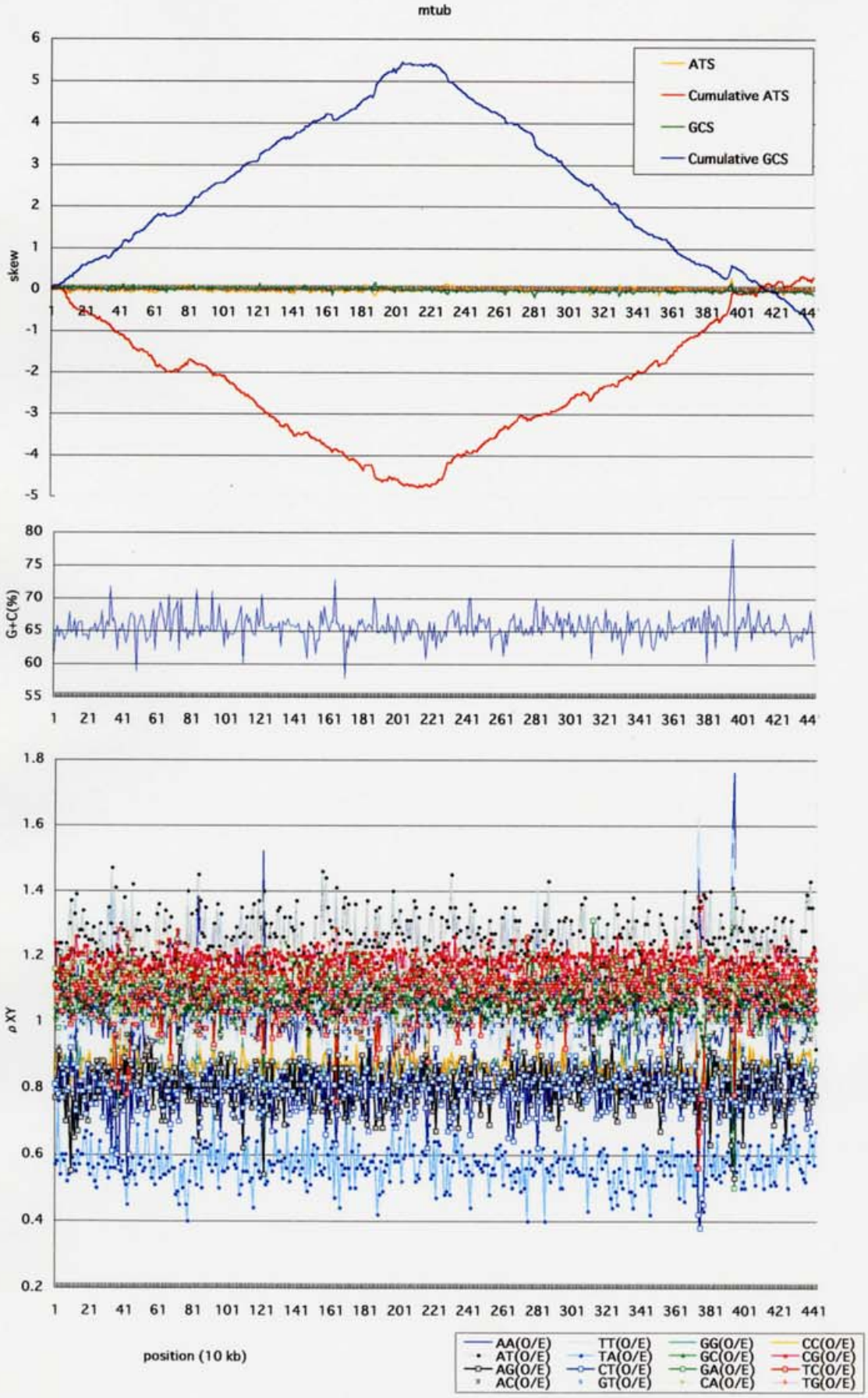


B-29 *Mycobacterium tuberculosis CDC1551* NC_002755 4,403,836 11 65.6

AE000516

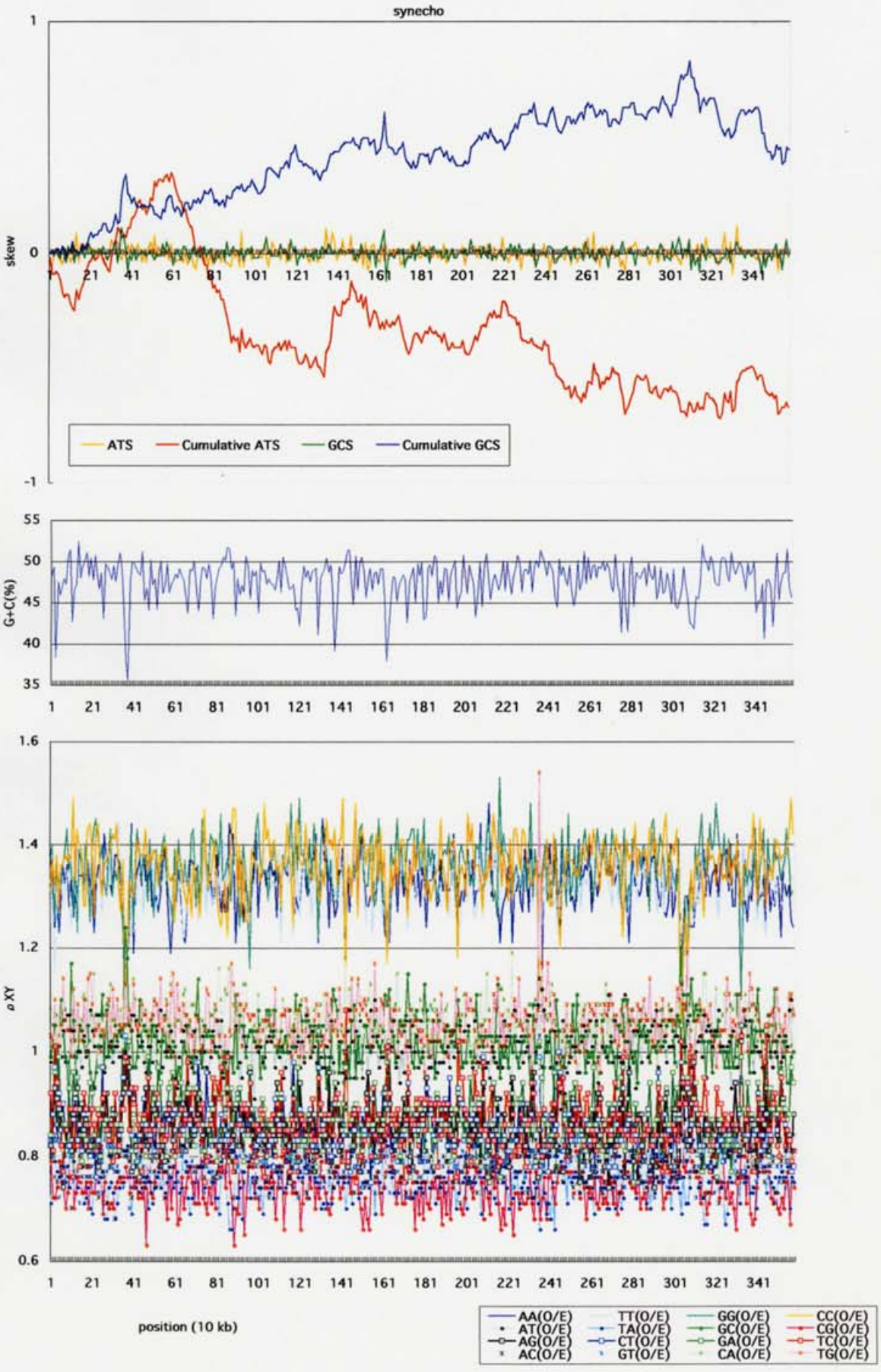


B-30 *Mycobacterium tuberculosis H37Rv* NC_000962 4,411,529 11 65.6



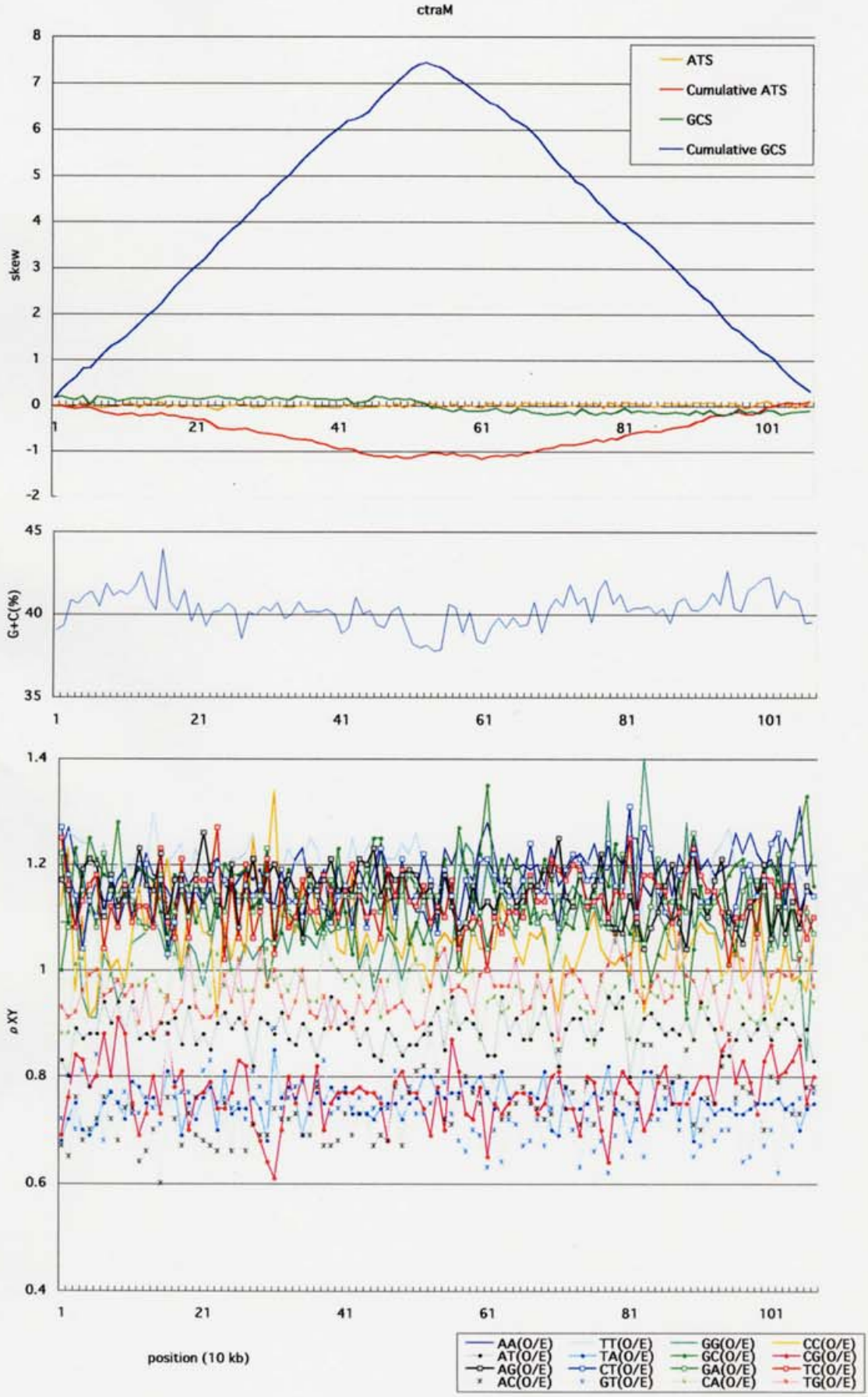
Bacteria; Cyanobacteria (1)

B-48 *Synechocystis sp. PCC 6803* NC_000911 3,573,470 11 47.7

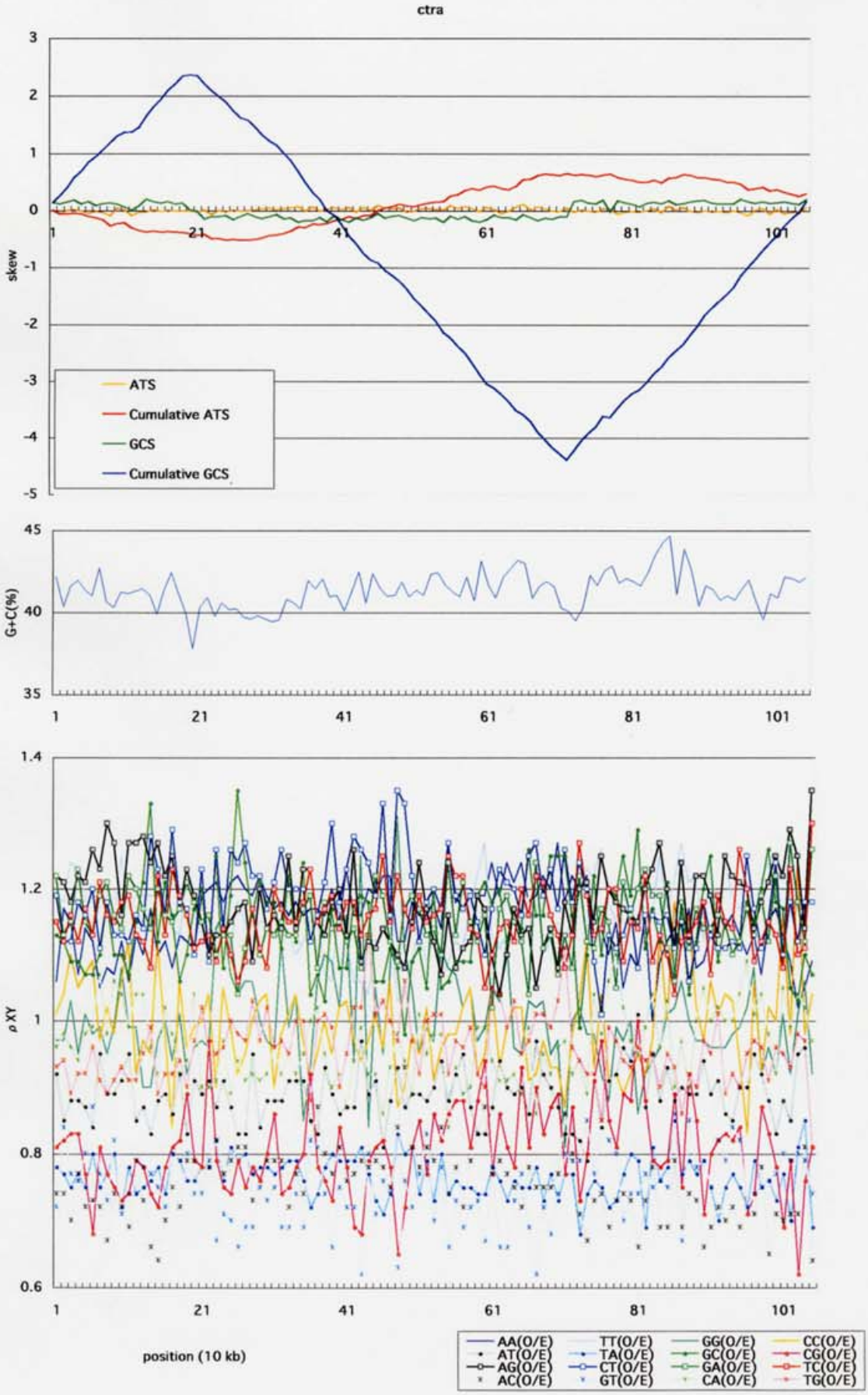


Bacteria; Chlamydiales (5)

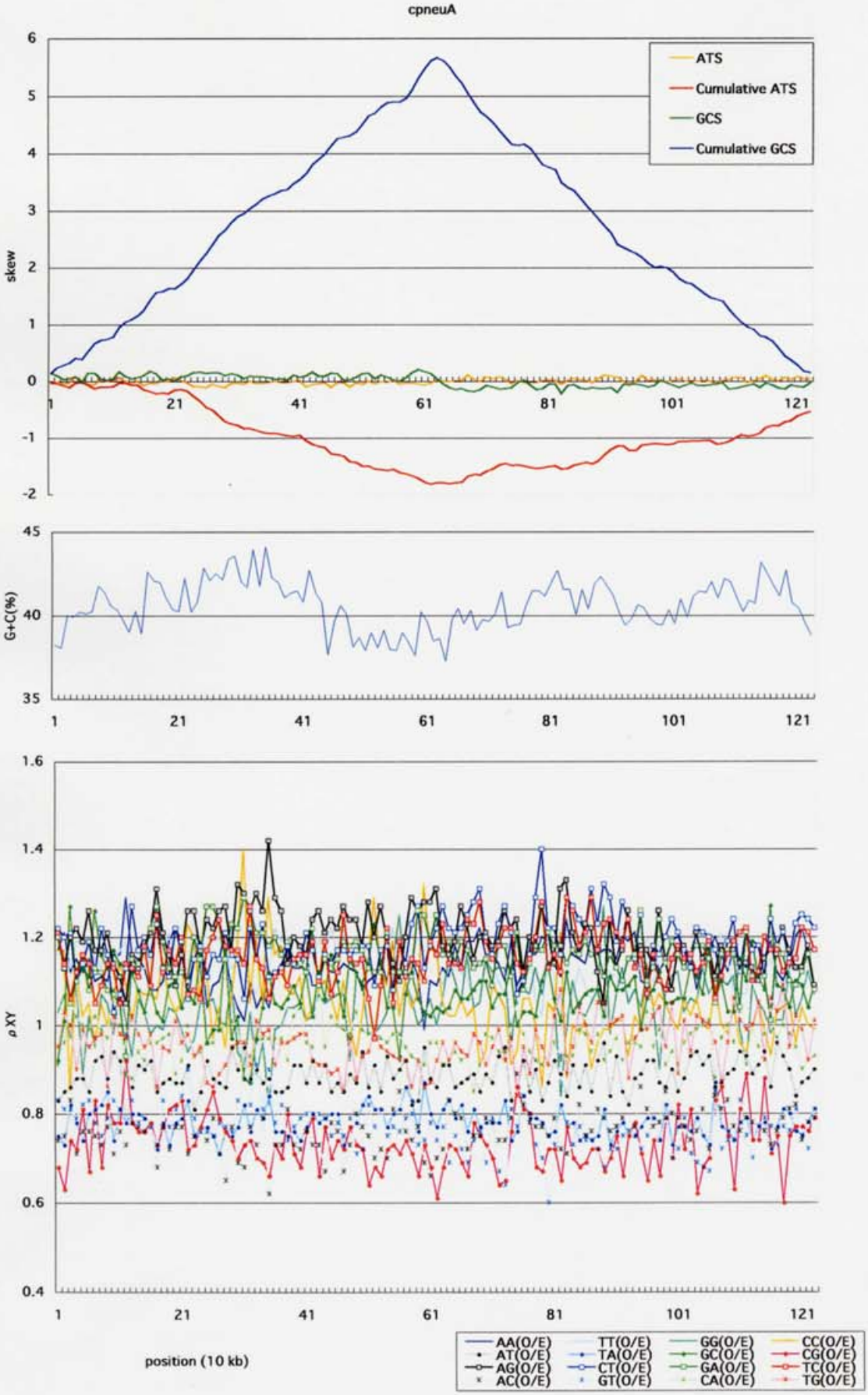
B-10 *Chlamydia muridarum* NC_002620 1,069,411 11 41.3



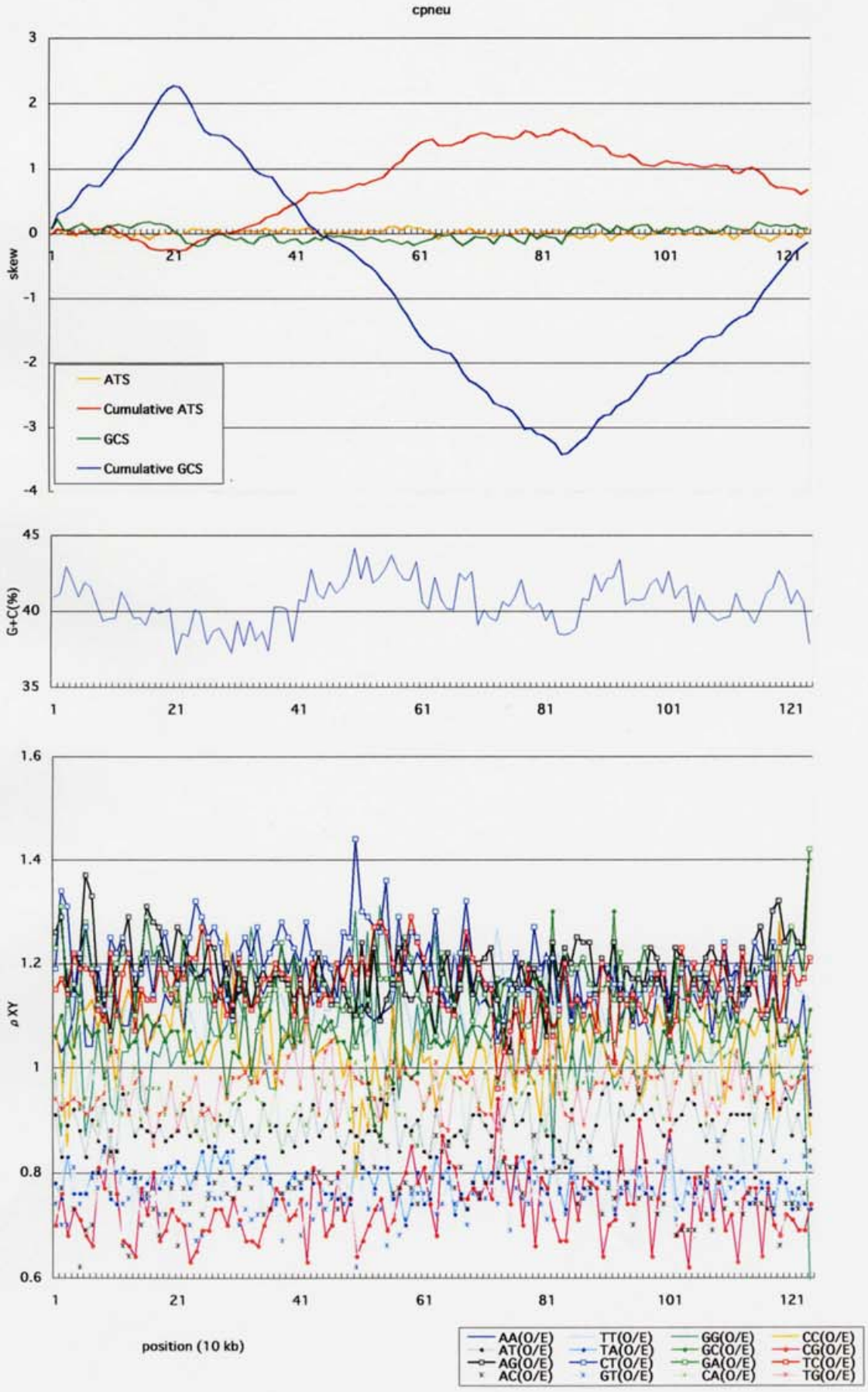
B-11 *Chlamydia trachomatis* NC_000117 1,042,519 11 40.3



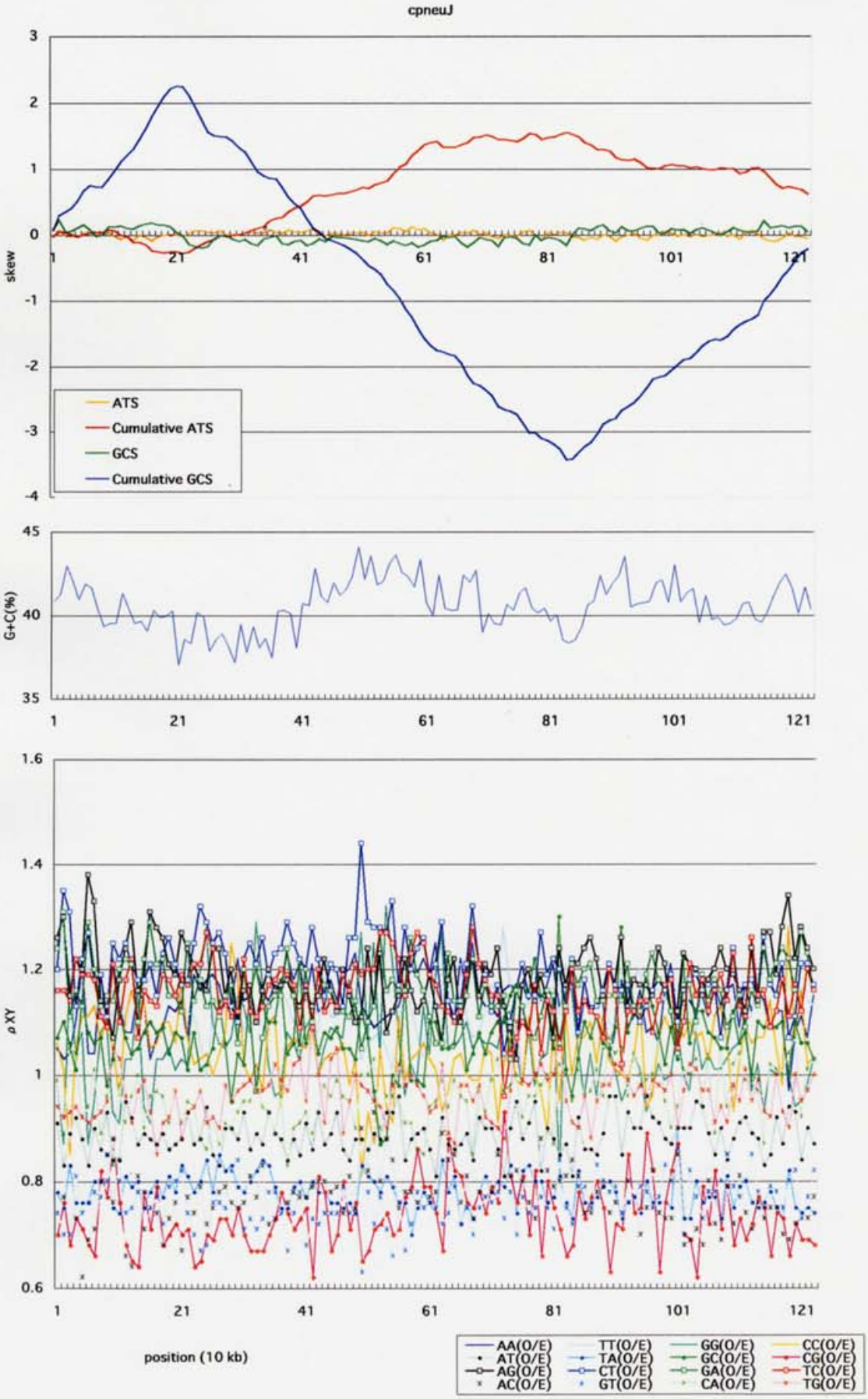
B−12 *Chlamydophila pneumoniae AR39* NC_002179 1,229,858 11 40.6



B-13 *Chlamydophila pneumoniae CWL029* NC_000922 1,230,230 11 40.6

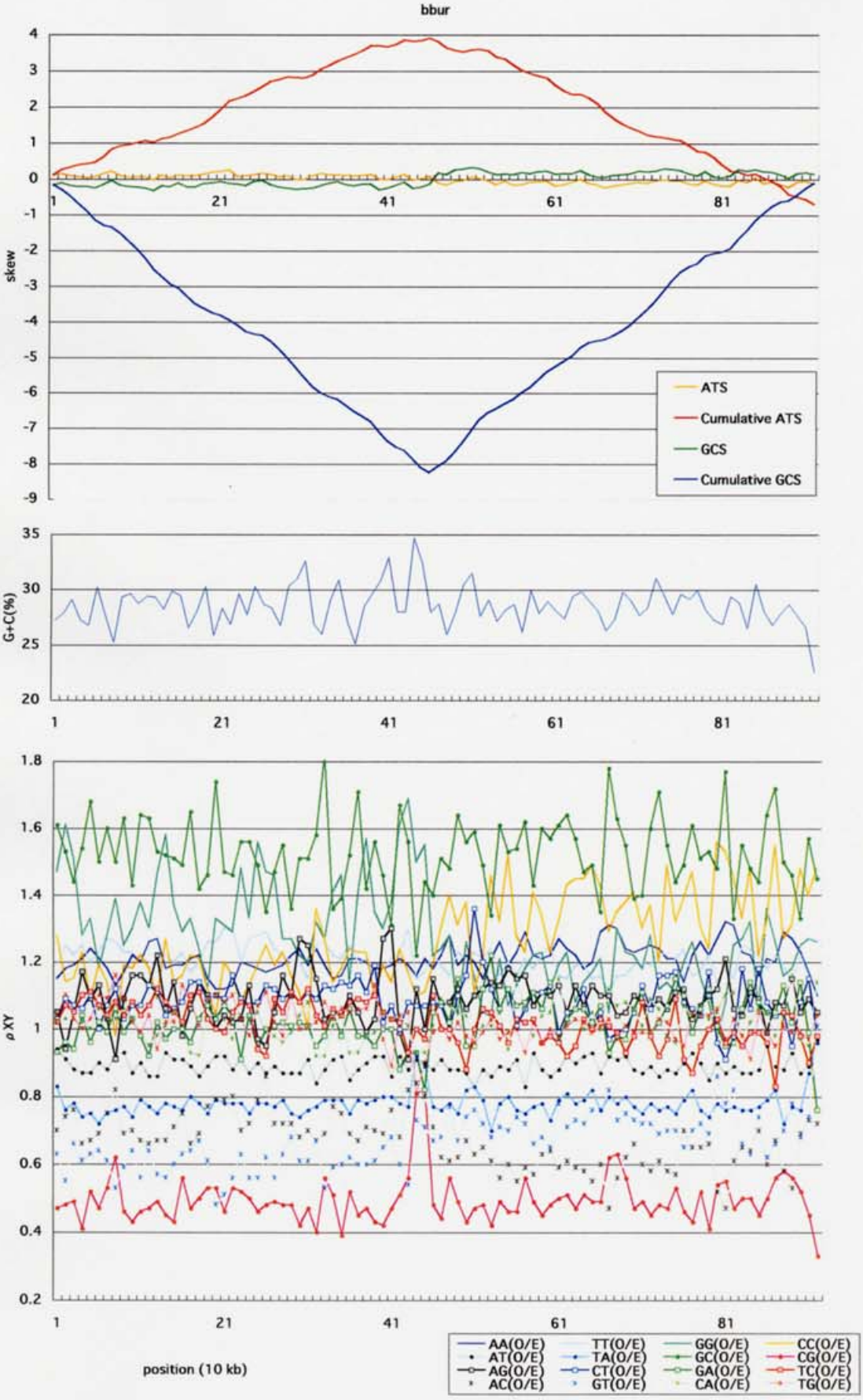


B-14 *Chlamydophila pneumoniae J138* NC_002491 1,226,565 11 40.6

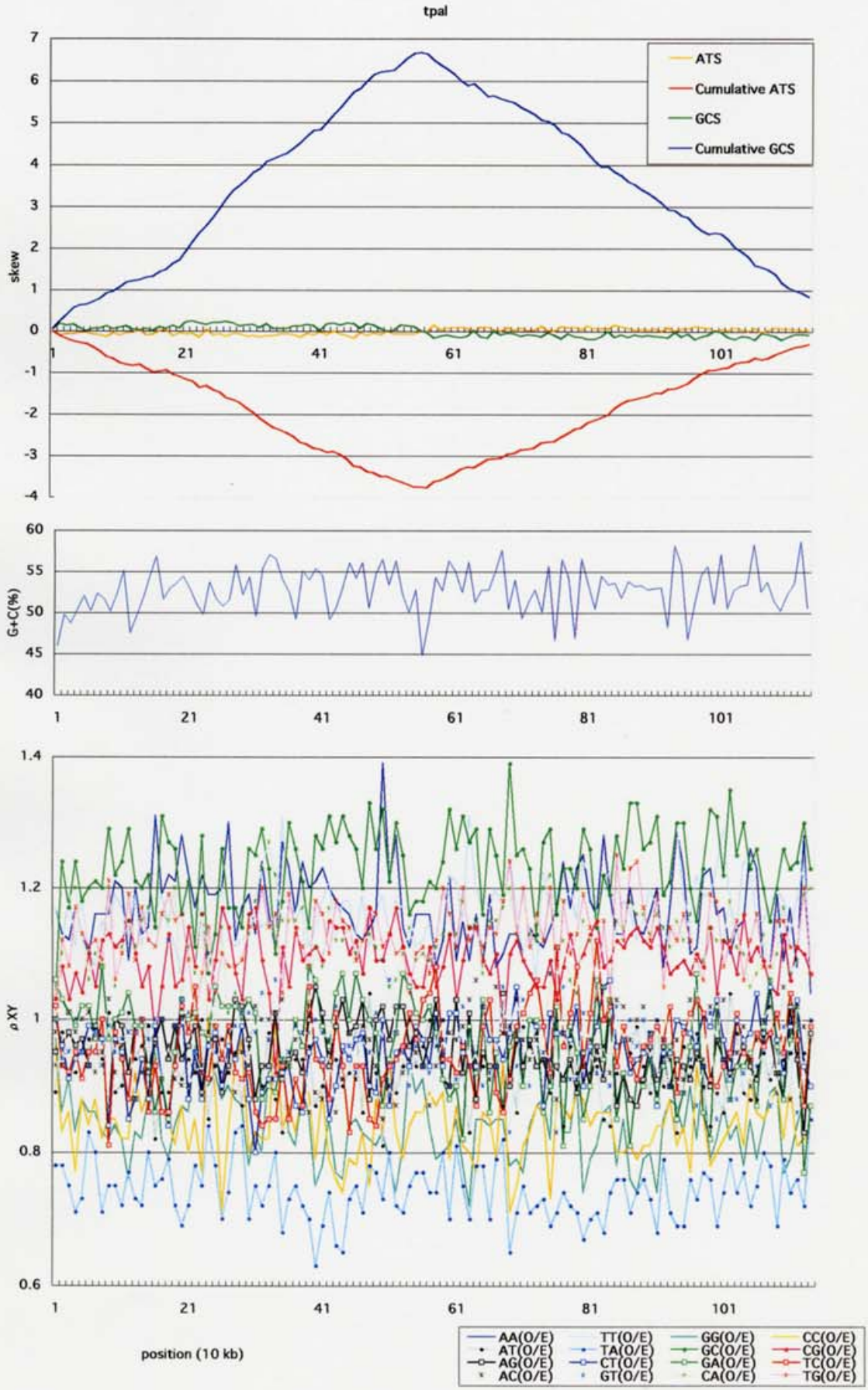


Bacteria; Spirochaetales (2)

B—6 *Borrelia burgdorferi* NC_001318 910,724 11 28.6

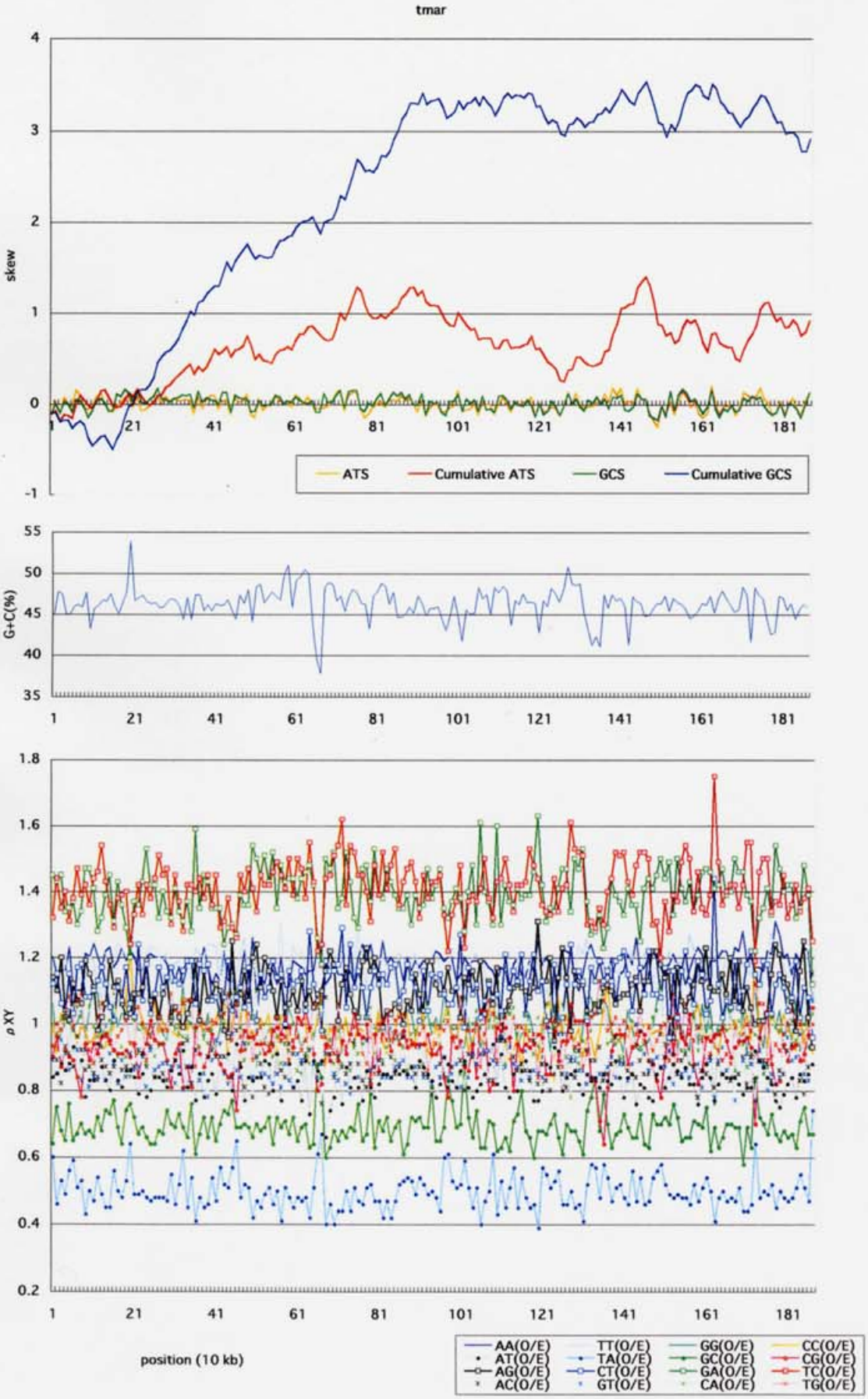


B-50 *Treponema pallidum* NC_000919 1,138,011 11 52.8



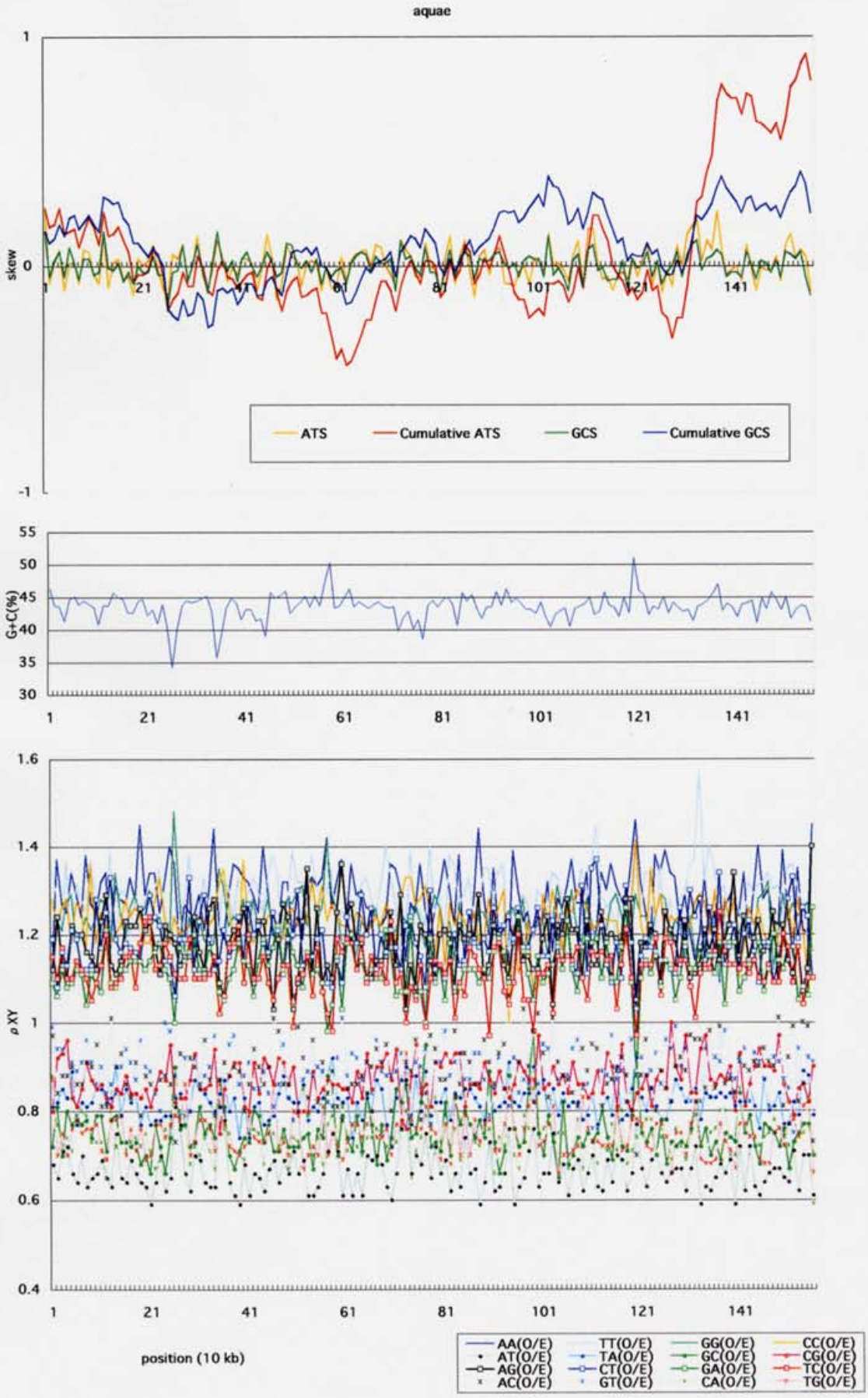
Bacteria; Thermotogales (1)

B-49 *Thermotoga maritima* NC_000853 1,860,725 11 46.3



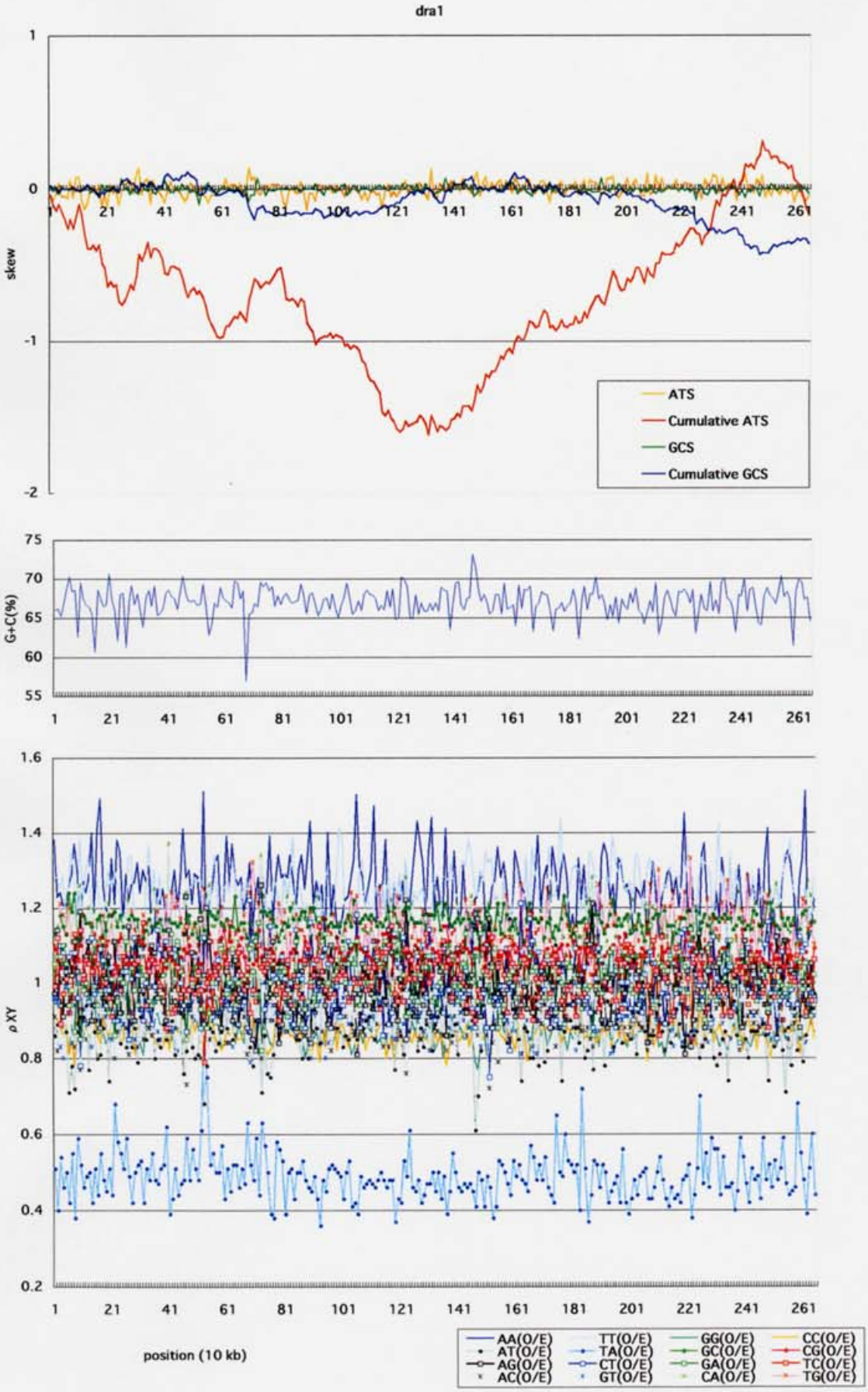
Bacteria; Aquificales (1)

B—3 *Aquifex aeolicus* NC_000918 1,551,335 11 43.5

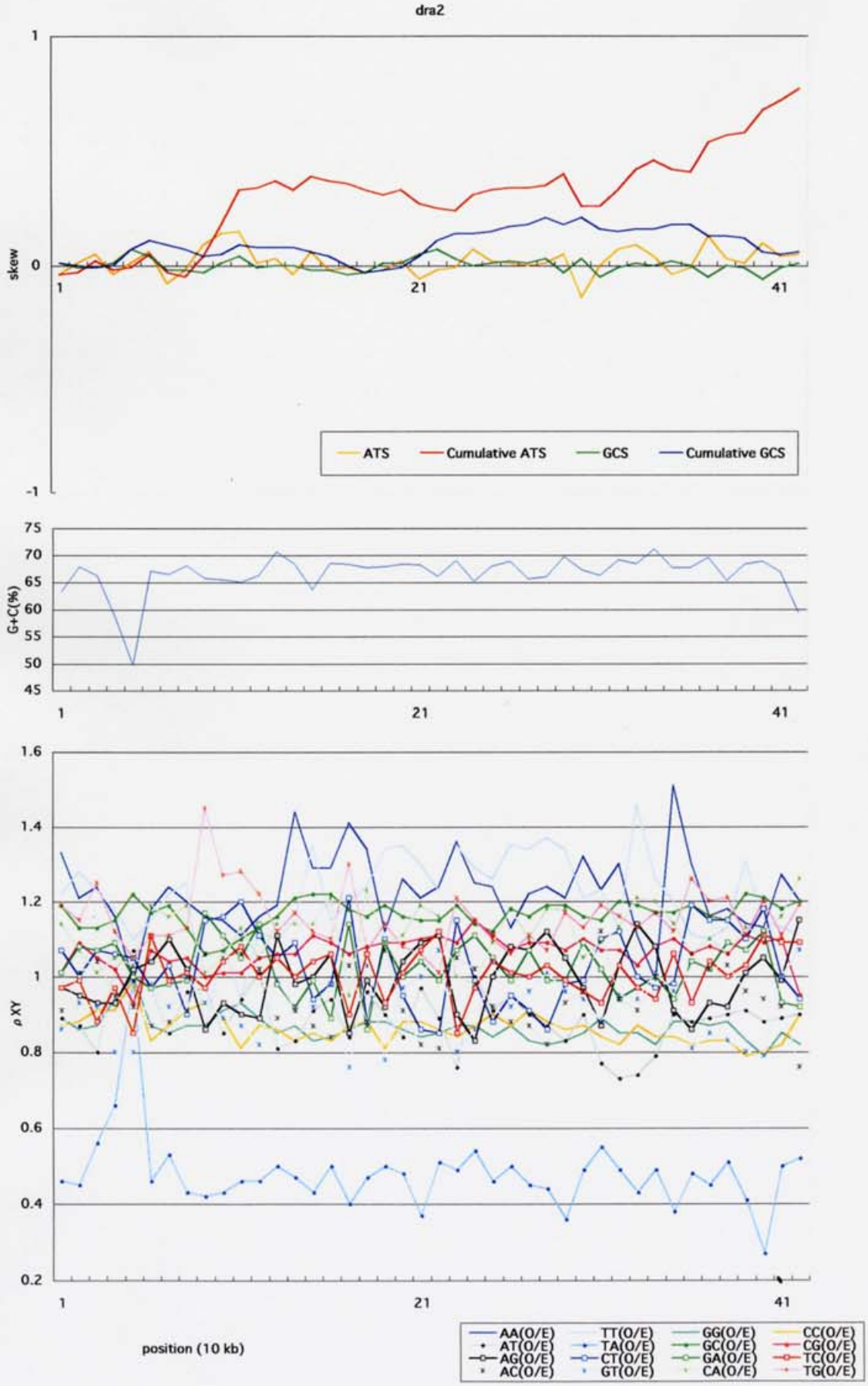


Bacteria; Thermus/Deinococcus group (2)

B-16 *Deinococcus radiodurans* chromosome1 NC_001263 2,648,638 11 67.0

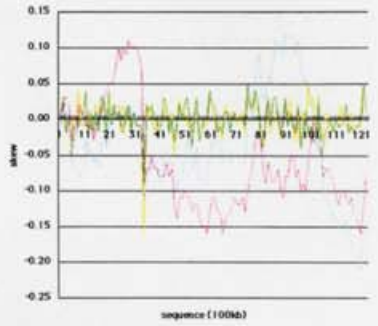


B-17 *Deinococcus radiodurans* chromosome2 NC_001264 412,348 11 66.7

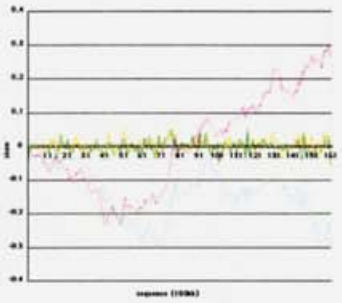


C. elegans

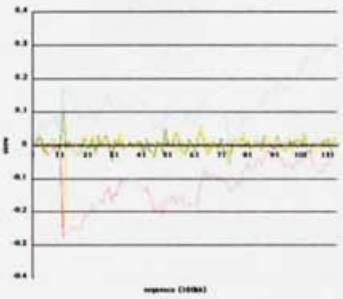
Chromosome I

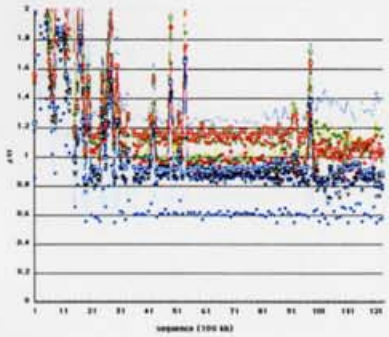


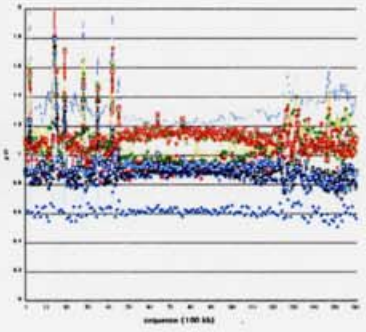
Chromosome II

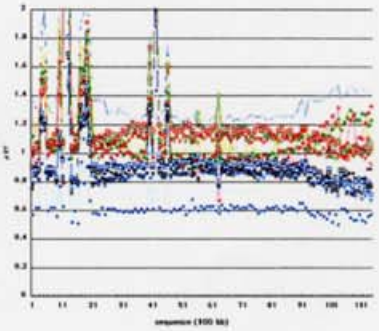


Chromosome III

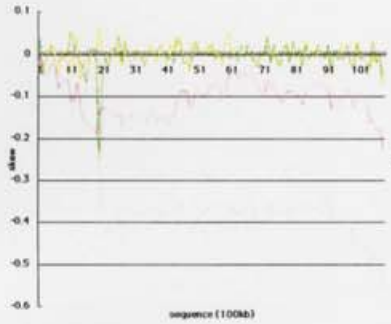




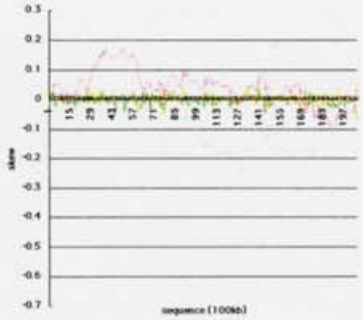




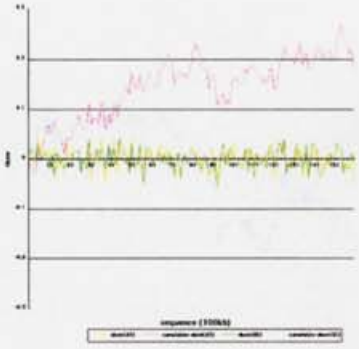
Chromosome IV

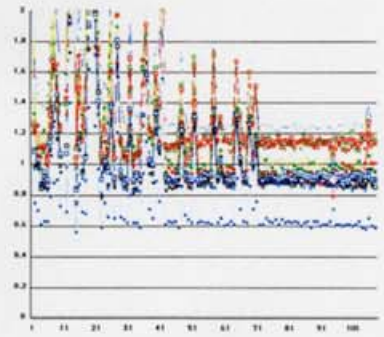


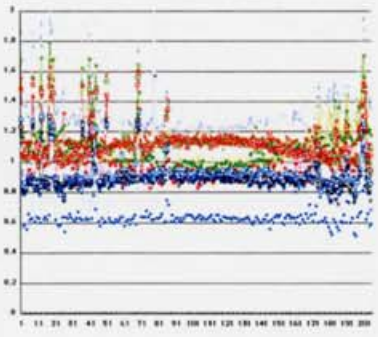
Chromosome V

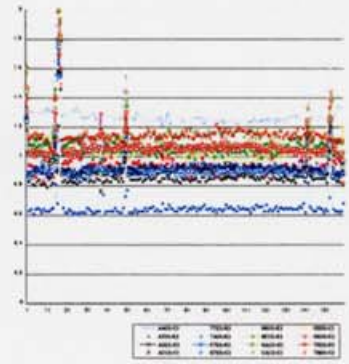


Chromosome X

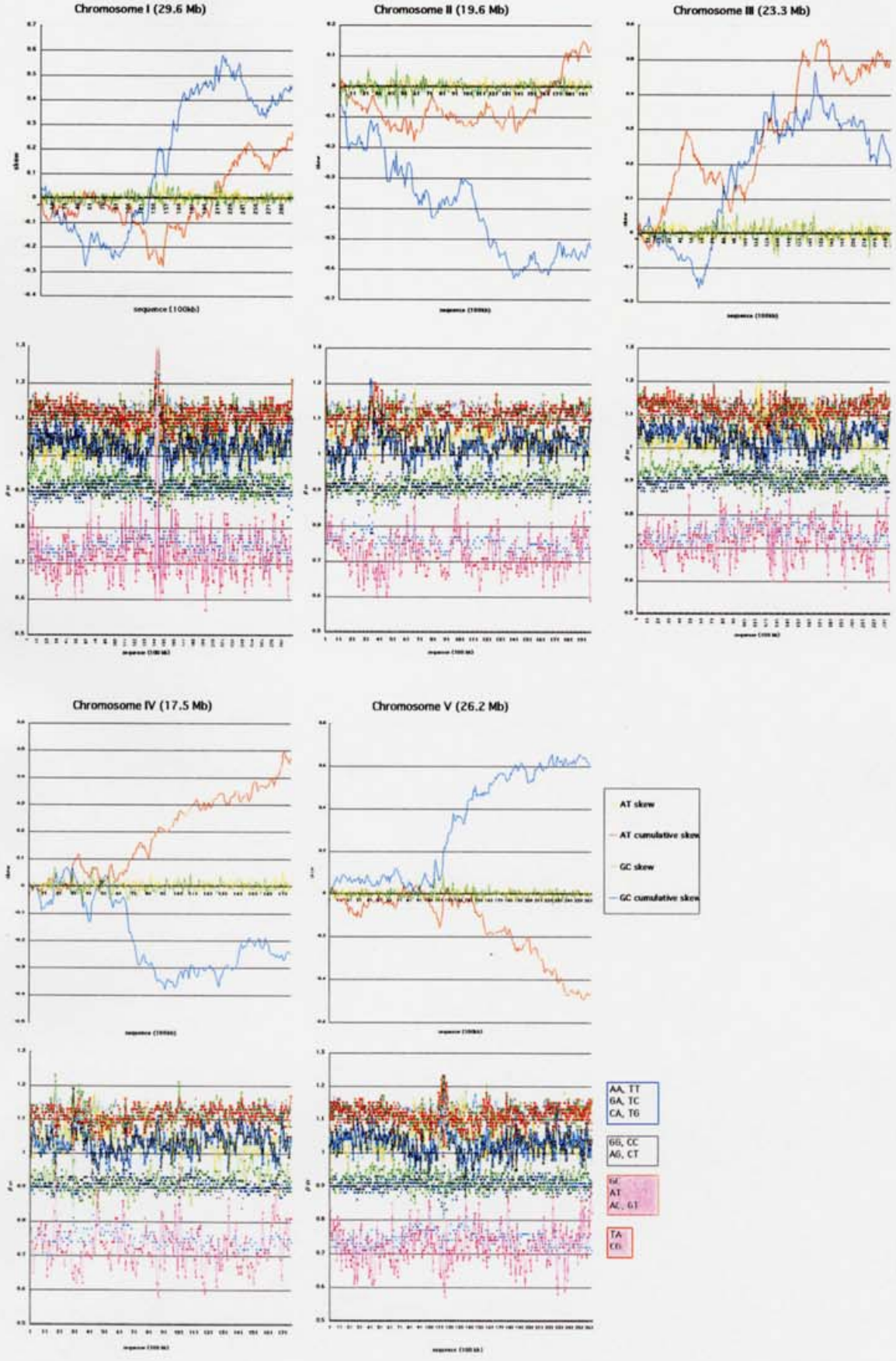


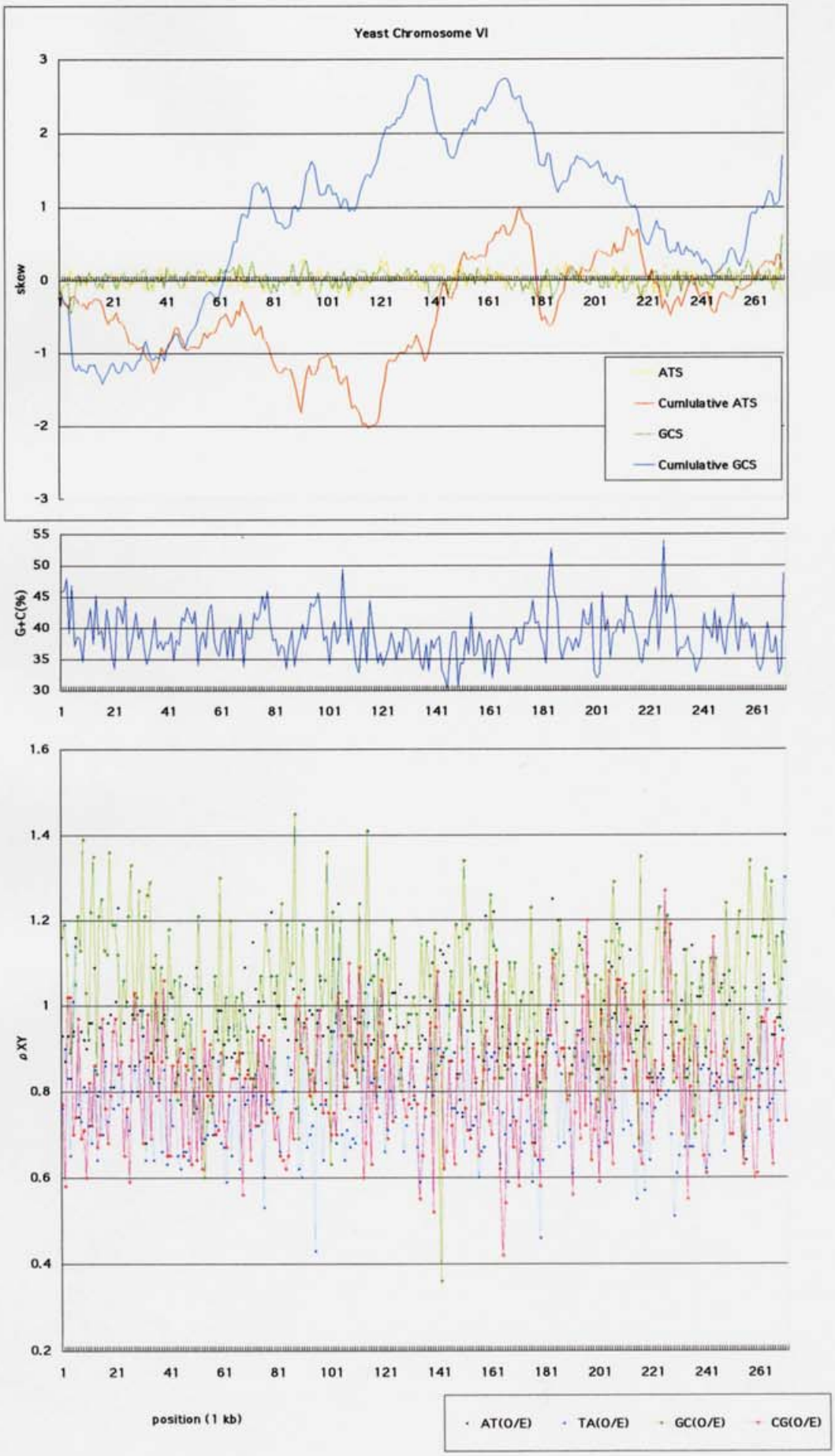






A.thaliana Genome (116 Mb)




Yeast Chromosome VI
skew
ATS
Cumulative ATS
GCS
Cumulative GCS
G+C(%)
ρ XY
position (1 kb)
AT(O/E)
TA(O/E)
GC(O/E)
CG(O/E)

J Mol Evol (2001) 53:364–376
DOI: 10.1007/s002390010226

JOURNAL OF MOLECULAR EVOLUTION

Skew of Mononucleotide Frequencies, Relative Abundance of Dinucleotides, and DNA Strand Asymmetry*

Chiharu Shioiri, Naoyuki Takahata

Department of Biosystems Science, Graduate University for Advanced Studies, Hayama, Kanagawa 240-0193, Japan

Received: 1 November 2000 / Accepted: 12 March 2001

Abstract. Based on 152 mitochondrial genomes and 36 bacterial chromosomes that have been completely sequenced, as well as three long contigs for human chromosomes 6, 21, and 22, we examined skews of mononucleotide frequencies and the relative abundance of dinucleotides in one DNA strand. Each group of these genomes has its own characteristics. Regarding mitochondrial genomes, both C_pG and G_pT are underrepresented, while either G_pG or C_pC or both are overrepresented. The relative frequency of nucleotide T vs A and of nucleotide G vs C is strongly skewed, due presumably to strand asymmetry in replication errors and unidirectional DNA replication from single origins. Exceptions are found in the plant and yeast mitochondrial genomes, each of which may replicate from multiple origins. Regarding bacterial genomes, the "universal" rule of C_pG deficiency is restricted to archaebacteria and some eubacteria. In other eubacteria, the most underrepresented dinucleotide is either T_pA or G_pT. In general, there are significant T vs A and G vs C skews in each half of the bacterial genome, although these are almost exactly canceled out over the whole genome. Regarding human chromosomes 6, 21, and 22, dinucleotide C_pG tends to be avoided. The relative frequency of mononucleotides exhibits conspicuous local skews, suggesting that each of these chromosomal segments contains more than one DNA replication origin. It is concluded that, when there are several replicons in a genomic region, not only the number of DNA replication origins but also the directionality is important and that the observed patterns of nucleotide frequencies in the genome strongly support the hypothesis of strand asymmetry in replication errors.

Key words: Replication origins — Nucleotide substitutions — T_pA/C_pG deficiency — T_pG/C_pT excess

Introduction

If nucleotide substitutions occur symmetrically in both leading and lagging strands of replicating nuclear DNA or both heavy and light strands of replicating mitochondrial (mt) DNA, the transition probability of one nucleotide to another has to have some symmetric relationships owing to the Watson–Crick pairing rule between the two strands. For example, if nucleotide A at a particular site is replaced by G, with probability P_{AG} in one strand and with Q_{AG} in the other strand, $Q_{AG} = P_{TC}$ and $P_{AG} = Q_{TC}$ are expected in each strand. Under the assumption of strand symmetry ($P_{AG} = Q_{AG}$), $P_{AG} = P_{TC}$ and $Q_{AG} = Q_{TC}$ must therefore hold true. Obviously, there are six such relationships in total. As a result, the frequency of A in one strand equals that of T in the same strand, and the same relationship is expected for the frequency of G and C.

Jukes and Cantor (1969) proposed a simple model in which all four nucleotides are substituted by each other with equal probability. It is remarkable that the proposal was made in a paper dealing with protein sequences, but in which the authors tried to infer the number of nucleo-

*This paper is dedicated to the late professor Thomas H. Jukes, whose contribution to molecular evolution was conceptually unprecedented.

Correspondence to: Dr. Naoyuki Takahata; *email:* takahata@soken.ac.jp

tide substitutions in light of the genetic code. Since then, a number of nucleotide substitution models were proposed to take into account realistic situations. Of these, Sueoka (1995) and Lobry (1995, 1996) considered the six-parameter model, which takes explicit account of strand symmetry or no-strand-bias conditions. As mentioned above, one immediate consequence of this sort of nucleotide substitution models is that at equilibrium, the frequency of A (f_A) is equal to that of T (f_T), and f_G is equal to f_C. In other words, there should not be any significant skew between f_A and f_T or between f_G and f_C. Without knowing any relevance to strand symmetry, Takahata and Kimura (1981) also proposed a five-parameter model and obtained formulas for multiple-hit corrections. This five-parameter model is exactly the same as the strand symmetry-based six-parameter model when the latter further assumes that the rate of nucleotide substitutions between A and T is equal to that between G and C.

However, it has been emphasized that the complementary strands differ from each other with respect to DNA replication, transcription, and recombination (Wu and Maeda 1987; Kunkel 1992; Waga and Stillman 1994; Beletskii and Bhagwat 1996; Francino and Ochman 1997; Freeman et al. 1998). Strand asymmetries in mutations, if present, may result in skewed nucleotide frequencies. This property has been of particular interest in attempts at identifying DNA replication origins (Grigoriev 1998; McLean et al. 1998; Lopez et al. 2000) and at examining the hypothesis of C-to-T deaminations in the nontranscribed strand (Beletskii and Bhagwat 1996). Also, it was suggested that possible disproportionate replication errors between the complementary strands allow the genome to increase genetic variance in a population, thereby increasing the efficacy of natural selection (Furusawa and Doi 1992, 1998).

On the other hand, S. Ohno and S. Karlin have conducted a series of examinations of oligonucleotide relative abundance. In particular, they noted a marked underrepresentation of C_pG and T_pA in eukaryote genes (Ohno 1988) and in a diverse set of prokaryotic, eukaryotic, organelle, and viral sequences (for a review see Karlin and Burge 1995; Karlin and Mrázek 1997). Karlin and Burge (1995) argued that the classical scenario based on methylation–deamination–mutation cannot account for the pervasive C_pG underrepresentations in all animal mitochondrial genomes. The stacking energy hypothesis was instead invoked in the context of flexibility for unwinding of the DNA double-helix.

The purpose of this paper is to examine skewed mononucleotide frequencies and relative abundance of dinucleotides simultaneously. Furthermore, since under strand symmetry, the dinucleotide frequency of C_pA, A_pC, C_pT, and T_pC is also expected to be equal to that of T_pG, G_pT, A_pG, and G_pA, respectively, we extend the analysis of skewness in mononucleotides to the case of dinucleotides. Based on completely sequenced mtDNAs and prokaryote genomes as well as long contigs for human and arabidopsis chromosomes, we show different patterns of mono- and dinucleotide frequencies in different genomes and provide evidence for strand asymmetry of nucleotide substitutions in relation to DNA replication origins.

Materials and Methods

We downloaded 152 complete mtDNA sequences (Table 1), 36 complete prokaryote chromosomes (Table 2), and several long contigs for human and *Arabidopsis thaliana* chromosomes, which were available as of October 2000. The total number of nucleotides analyzed is more than 200 megabases (Mb). In a given DNA sequence, we measure skewed mononucleotide frequencies by

$$\mathrm{ATS} = \frac{f_A - f_T}{f_A + f_T} \quad \text{and} \quad \mathrm{GCS} = \frac{f_G - f_C}{f_G = f_C} \tag{1}$$

and relative abundance of dinucleotides by

$$\rho_{XY} = \frac{f_{XY}}{f_X f_Y} \tag{2}$$

where f_{XY} denotes the observed frequency of dinucleotide X_pY and f_X denotes the observed frequency of nucleotide X. In Eq. (2), we do not use the symmetrized version (ρ^*_{XY}) proposed by Burge et al. (1992), because our concern is in part with the presence or absence of strand asymmetries. Although the frequency of a dinucleotide happens to be close to that of the inverted complementary dinucleotide, such a relationship is definitely violated in some genomes. For instance, f_{AG} = 4.8% is much lower than f_{CT} = 8.7% in the human mtDNA, although f_{AG} and f_{CT} are almost the same in the prokaryote genome.

McLean et al. (1998) examined ATS and GCS at the third codon positions only, in completely sequenced 12 prokaryote genomes. This was based on the consideration that the base composition skew at the third codon position reflects mutational biases more faithfully than that over all bases. While this approach is sensible and certainly feasible for mtDNA and prokaryote genomes, it is not practically so in eukaryote genomes, in which the vast majority are noncoding and not all genes are identified. For this reason, we do not distinguish between coding and noncoding regions. Rather, examining a large number of genomes, we try to identify which sequences or genomes in each group are atypical in terms of skew and relative abundance of mono- and dinucleotides.

Results

mtDNAs

Among the complete mtDNA sequences deposited in GenBank, 120 come from animals, 5 from fungi, 10 from plants, and 14 from protozoans (Table 1). The jawed vertebrate mtDNAs are all similar in GC content (ca. 40%) and identical in genetic code (Jukes and Osawa 1990, 1993; Osawa 1994). In other organisms, the GC content varies greatly; it is lowest in *Apis mellifera ligustica* (15.1%) and highest in *Balanoglossus camosus*

Table 1. List of 152 species in which the mitochondrial genome is completely sequenced[a]

Species	bp	G + C[b]	G[c]	ATS[d]	GCS[d]
1. *Homo sapiens*	16,569	44.4	2	0.11	–0.41
2. *Pan paniscus*	16,563	43.4	2	0.11	–0.41
3. *Pan troglodytes*	16,554	43.7	2	0.11	–0.41
4. *Gorilla gorilla*	16,364	43.9	2	0.10	–0.40
5. *Pongo pygmaeus*	16,389	45.7	2	0.12	–0.42
6. *Pongo pygmaeus abelii*	16,499	45.9	2	0.13	–0.43
7. *Hylobates lar*	16,472	45.5	2	0.12	–0.40
8. *Papio hamadryas*	16,521	43.7	2	0.12	–0.40
9. *Tupaia belangeri*	16,754	40.8	2	0.10	–0.29
10. *Dasypus novemcinctus*	17,056	38.9	2	0.13	–0.34
11. *Erinaceus europaeus*	17,447	32.6	2	0.01	–0.23
12. *Talpa europaea*	16,884	38.9	2	0.12	–0.26
13. *Oryctolagus cuniculus*	17,245	40.2	2	0.05	–0.32
14. *Sciurus vulgaris*	16,507	37.0	2	0.02	–0.32
15. *Mus musculus*	16,295	36.7	2	0.09	–0.33
16. *Rattus norvegicus*	16,300	38.7	2	0.11	–0.36
17. *Myoxus glis*	16,602	36.2	2	0.02	–0.30
18. *Cavia porcellus*	16,801	39.3	2	0.06	–0.26
19. *Artibeus jamaicensis*	16,651	37.9	2	0.04	–0.31
20. *Loxodonta africana*	16,866	38.8	2	0.07	–0.30
21. *Equus asinus*	16,670	42.1	2	0.12	–0.37
22. *Equus caballus*	16,660	42.0	2	0.11	–0.36
23. *Ceratotherium simum*	16,832	40.9	2	0.13	–0.37
24. *Rhinoceros unicornis*	16,829	40.2	2	0.12	–0.37
25. *Orycteropus afer*	16,816	38.1	2	0.07	–0.34
26. *Canis familiaris*	16,728	39.7	2	0.05	–0.29
27. *Felis catus*	17,009	40.3	2	0.09	–0.30
28. *Halichoerus grypus*	16,797	41.7	2	0.13	–0.32
29. *Phoca vitulina*	16,826	41.7	2	0.13	–0.32
30. *Bos taurus*	16,338	39.4	2	0.10	–0.32
31. *Ovis aries*	16,616	38.9	2	0.10	–0.33
32. *Sus scrofa*	16,613	39.5	2	0.15	–0.33
33. *Lama pacos*	16,652	40.9	2	0.08	–0.29
34. *Hippopotamus amphibius*	16,407	42.6	2	0.14	–0.34
35. *Balaenoptera musculus*	16,402	40.7	2	0.10	–0.36
36. *Balaenoptera physalus*	16,398	40.6	2	0.10	–0.34
37. *Physeter catodon*	16,428	43.1	2	0.12	–0.38
38. *Macropus robustus*	16,896	39.2	2	0.09	–0.34
39. *Didelphis virginiana*	17,084	33.2	2	0.06	–0.27
40. *Ornithorhynchus anatinus*	17,019	37.1	2	0	–0.27
41. *Corvus frugilegus*	16,932	44.3	2	0.10	–0.34
42. *Smithornis sharpei*	17,344	45.2	2	0.10	–0.44
43. *Vidua chalybeata*	16,895	45.8	2	0.15	–0.35
44. *Falco peregrinus*	18,068	44.4	2	0.18	–0.39
45. *Gallus gallus*	16,775	46.0	2	0.12	–0.41
46. *Aythya americana*	16,616	48.4	2	0.14	–0.35
47. *Ciconia boyciana*	17,622	46.3	2	0.15	–0.38
48. *Ciconia ciconia*	17,347	46.3	2	0.14	–0.38
49. *Rhea americana*	16,714	46.9	2	0.08	–0.37
50. *Struthio camelus*	16,591	44.7	2	0.10	–0.36
51. *Alligator mississippiensis*	16,646	43.0	2	0.10	–0.37
52. *Dinodon semicarinatus*	17,191	39.9	2	0.16	–0.39
53. *Eumeces egregius*	17,407	44.2	2	0.09	–0.30
54. *Chelonia mydas*	16,497	39.5	2	0.17	–0.39
55. *Chrysemys picta*	16,866	38.8	2	0.13	–0.34
56. *Pelomedusa subrufa*	16,787	38.7	2	0.11	–0.37
57. *Xenopus laevis*	17,553	37.0	2	0.05	–0.27
58. *Typhlonectes natans*	17,005	45.1	2	0.10	–0.29
59. *Carassius auratus*	16,578	42.6	2	0.09	–0.24
60. *Crossostoma lacustre*	16,558	45.5	2	0.08	–0.26
61. *Cyprinus carpio*	16,575	43.3	2	0.12	–0.27
62. *Danio rerio*	16,890	39.8	2	0.06	–0.20
63. *Oncorhynchus mykiss*	16,642	46.0	2	0.03	–0.26
64. *Salmo salar*	16,665	45.2	2	0.04	–0.28
65. *Salvelinus alpinus*	16,659	45.5	2	0.03	–0.25
66. *Salvelinus fontinalis*	16,624	45.2	2	0.03	–0.25
67. *Gadus morhua*	16,696	42.4	2	–0.03	–0.21
68. *Paralichthys olivaceus*	17,090	46.5	2	0.02	–0.28
69. *Polypterus ornatipinnis*	16,624	39.8	2	0.07	–0.29
70. *Protopterus dolloi*	16,646	42.2	2	0	–0.25
71. *Latimeria chalumnae*	16,407	41.7	2	0.18	–0.28
72. *Mustelus manazo*	16,707	38.3	2	0	–0.27
73. *Scyliorhinus canicula*	16,697	38.0	2	–0.01	–0.26
74. *Squalus acanthias*	16,738	39.8	2	0.01	–0.26
75. *Raja radiata*	16,783	40.3	2	0.02	–0.29
76. *Lampetra fluviatilis*	16,159	38.6	2	0.03	–0.26
77. *Petromyzon marinus*	16,201	37.3	2	0.03	–0.28
78. *Branchiostoma floridae*	15,083	37.3	5	–0.14	0.15
79. *Branchiostoma lanceolatum*	15,076	37.4	5	–0.14	0.15
80. *Halocynthia roretzi*	14,771	31.7	13	–0.29	0.46
81. *Balanoglossus carnosus*	15,708	48.6	9	–0.03	–0.30
82. *Laqueus rubellus*	14,017	41.6	5	–0.29	0.27
83. *Terebratulina retusa*	15,451	42.8	5	0.03	–0.29
84. *Arbacia lixula*	15,719	37.5	9	–0.06	–0.09
85. *Paracentrotus lividus*	15,696	39.7	9	0.02	–0.13
86. *Strongylocentrotus purpuratus*	15,650	41.0	9	–0.03	–0.10
87. *Asterina pectinifera*	16,260	38.7	9	0.06	–0.27
88. *Florometra serratissima*	16,005	27.2	9	–0.27	0.15
89. *Apis mellifera ligustica*	16,343	15.1	5	0.02	–0.27
90. *Bombyx mori*	15,643	18.7	5	0.06	–0.22
91. *Ceratitis capitata*	15,980	22.5	5	0.02	–0.19
92. *Drosophila melanogaster*	19,517	17.8	5	0.02	–0.15
93. *Drosophila yakuba*	16,019	21.4	5	0.01	–0.14
94. *Anopheles gambiae*	15,363	22.4	5	0.03	–0.15
95. *Anopheles quadrimaculatus*	15,455	22.6	5	0.04	–0.18
96. *Locusta migratoria*	15,722	24.7	5	0.18	–0.18
97. *Artemia franciscana*	15,822	35.6	5	–0.04	–0.01
98. *Daphnia pulex*	15,333	37.7	5	0.01	–0.12
99. *Penaeus monodon*	15,984	29.4	5	0	–0.14
100. *Ixodes hexagonus*	14,539	27.3	5	0.03	–0.37
101. *Rhipicephalus sanguineus*	14,710	22.0	5	–0.03	–0.10
102. *Lumbricus terrestris*	14,998	38.4	5	–0.03	–0.18
103. *Platynereis dumerilii*	15,619	35.9	5	–0.03	–0.14
104. *Albinaria coerulea*	14,130	29.4	5	–0.07	0.06
105. *Cepaea nemoralis*	14,100	40.2	5	–0.12	0.06
106. *Pupa strigosa*	14,189	38.9	5	–0.10	0.06
107. *Crassostrea gigas*	18,224	36.7	5	–0.13	0.20
108. *Katharina tunicata*	15,532	30.5	5	–0.10	0.22
109. *Loligo bleekeri*	17,211	28.7	5	0.09	–0.36
110. *Ascaris suum*	14,284	28.0	5	–0.38	0.45
111. *Caenorhabditis elegans*	13,794	23.8	5	–0.18	0.25
112. *Onchocerca volvulus*	13,747	26.7	5	–0.47	0.49
113. *Fasciola hepatica*	14,462	37.8	5	–0.48	0.47
114. *Paragonimus westermani*	14,964	48.3	14	–0.39	0.29
115. *Schistosoma japonicum*	14,085	29.0	5	–0.30	0.42
116. *Schistosoma mansoni*	14,415	31.5	5	–0.26	0.46
117. *Schistosoma mekongi*	14,072	27.8	5	–0.28	0.48

Table 1. Continued

Species	bp	G + C[b]	G[c]	ATS[d]	GCS[d]
118. *Echinococcus multilocularis*	13,738	31.0	14	−0.40	0.51
119. *Taenia crassiceps*	13,503	26.0	5	−0.31	0.41
120. *Metridium senile*	17,443	38.1	4	−0.13	0.11
121. *Pichia canadensis*	27,694	18.1	4	0.02	0.12
122. *Saccharomyces cerevisiae*	85,779	17.1	3	0.02	0.06
123. *Podospora anserina*	100,314	30.1	4	0.02	0.11
124. *Schizosaccharomyces pombe*	19,431	30.1	4	−0.03	0.05
125. *Allomyces macrogynus*	57,473	39.5	4	−0.04	0.06
126. *Arabidopsis thaliana*	366,923	44.8	1	0.01	−0.01
127. *Beta vulgaris* var. *altissima*	368,799	43.9	1	0	0
128. *Marchantia polymorpha*	186,609	42.4	1	−0.01	0.01
129. *Chlamydomonas eugametos*	22,897	34.6	1	0.01	0.15
130. *Chlamydomonas reinhardtii*	15,758	45.2	1	0.01	0.01
131. *Scenedesmus obliquus*	42,781	36.2	22	−0.04	0.12
132. *Pedinomas minor*	25,137	22.2	4	−0.19	0.13
133. *Prototheca wickerhamii*	55,328	25.8	1	0.03	−0.02
134. *Rhodomonas salina*	48,063	29.8	1	0.05	0.06
135. *Chondrus crispus*	25,836	27.9	4	0.05	0.04
136. *Cyanidioschyzon merolae*	32,211	27.1	1	0.01	0.06
137. *Porphyra purpurea*	36,753	33.5	4	0.07	0.04
138. *Paramecium aurelia*	40,469	41.2	4	0.14	0.06
139. *Plasmodium falciparum*	5,967	31.6	4	−0.05	0.01
140. *Plasmodium reichenowi*	5,966	31.7	4	−0.05	0.01
141. *Tetrahymena pyriformis*	47,296	21.3	4	−0.04	0.02
142. *Acanthamoeba castellanii*	41,591	29.4	4	−0.09	0.11
143. *Naegleria gruberi*	49,843	22.2	1	−0.09	0.17
144. *Leishmania tarentolae*	20,992	21.1	1	−0.03	0.10
145. *Physarum polycephalum*	65,862	25.9	1	0.03	0.02
146. *Dictyostelium discoideum*	55,564	27.4	1	0.20	0.24
147. *Malawimonas jakobiformis*	47,328	26.1	1	0.03	0
148. *Reclinomonas americana*	69,034	26.1	1	0	0.13
149. *Cafeteria roenbergensis*	43,159	27.3	4	−0.05	0.11
150. *Chrysodidymus synuroideus*	34,119	24.1	1	0.02	0.05
151. *Ochromonas danica*	41,035	26.2	1	0.06	0.01
152. *Phytophthora infestans*	37,957	22.3	1	0	0.05

[a] The sequences are grouped into nine: mammals (1 to 40), aves (41 to 50), reptiles (51 to 56), amphibians (57 and 58), fish (59 to 80) including cephalochordata and urochordata, invertebrates (81 to 120) including hemichordata, fungi (121 to 125), plants (126 to 137), and protozoa (138 to 152).

[b] GC content (%).

[c] Genetic code number adopted in NCBI is based primarily on Jukes and Osawa (1993): 1, stands for the standard code; 2, for the vertebrate; 3, for the yeast; 4, for the protozoan and mycoplasma; 5, for the invertebrate; 9, for the echinoderm; 13, for the ascidian; 14, for the flatworm; 22, for *S. obliquus* mitochondrial code.

[d] Skews between A and T and between G and C, defined in formula (1).

(48.6%). The genetic code is also at variance, and the code numbers used in NCBI are based primarily on Jukes and Osawa (1993).

In most organisms, excluding cephalochordata and some invertebrates and plants, ATS or GCS over the whole mitochondrial genome often deviates significantly from 0, and the absolute ATS and GCS values differ from each other (Table 1). If we compute the cumulative value of ATS or GCS over consecutive small windows, it increases or decreases monotonously. There are two patterns showing either ATS > 0 and GCS < 0 or ATS < 0 and GCS > 0. However, it should be noted that if ATS or GCS > 0 in one strand, necessarily ATS or GCS < 0 in the other stand. Accordingly, unless we compare the same strand such as the heavy strand, the distinction between the two types does not have any biological meaning. In any event, significantly skewed ATS or GCS implies that the tempo and mode in replication errors may differ between the complementary strands. Although neither strand of mtDNA is replicated discontinuously, the bias results from the particularities of mtDNA replication, and it is less likely that transcription can bias the occurrence of nucleotide substitutions between the two strands (Francino and Ochman 1997).

No significant skew is observed in the whole mtDNAs of three embryophyta (|ATS| and |GCS| $\leq$0.01). Thus, whereas strand asymmetry is the rule rather than the exception in animal mtDNAs, it is erased in these plant mtDNAs. The size of plant mtDNAs is generally 10 to 20 times larger than that of animal mtDNAs. Because of this large genome size, the DNA molecule may replicate from multiple origins, although the molecular cause for unbiased patterns of nucleotide frequencies await further scrutiny. Similarly, the yeast (*Saccharomyces cerevisiae*) mtDNA is relatively long, the genome contains three replication origins as well as four other replication origin-like positions, and replication might be bidirectional (Lecrenier and Foury 2000). Nonetheless, both ATS and GCS values over the genome tend to be positive, and strand symmetry is not recovered. This suggests that, in addition to the number of replication origins, information on the mode and direction of DNA replication is definitely important. In fact, even when there are multiple replication origins on the genome, strand asymmetry can still be expected if the DNA molecule replicates in the same direction. The proposed mechanism of the yeast mtDNA replication system is still hypothetical. Phenomenologically, the system is similar to that of animal mtDNAs, but different from that of plant mtDNAs.

In mtDNA, there are several characteristic dinucleotide profiles. First, in accord with ATS $\neq$ 0 and GCS $\neq$ 0, f_{XY} is quite different from the frequency of the inverted complementary dinucleotide. In the jawed vertebrate mtDNAs, $f_{CA} > f_{TG}$, $f_{AC} > f_{GT}$, $f_{TC} > f_{GA}$, and $f_{CT} > f_{AG}$ are observed, but in other mtDNAs these relation-

Table 2. List of 36 bacterial chromosomes[a]

Species	bp	G + C	G	ATS	GCS
1. *Aeropyrum pernix*	1,669,695	56.3	11	−0.01	−0.01
2. *Archaeoglobus fulgidus*	2,178,400	48.6	11	0	0
3. *Methanobacterium thermoautotrophicum* delta H	1,751,377	49.5	11	−0.01	0
4. *Methanococcus jannaschii*	1,664,970	31.4	11	0	0.01
5. *Pyrococcus abyssi*	1,765,118	44.7	11	0	0
6. *Pyrococcus horikoshii* OT3	1,738,505	41.9	11	0	−0.01
7. *Aquifex aeolicus*	1,551,335	43.5	11	0.01	0
8. *Bacillus subtilis*	4,214,814	43.5	11	0	0
9. *Borrelia burgdorferi*	910,724	28.6	11	−0.01	0
10. *Buchnera* sp. APS	640,681	26.3	11	0.01	0.01
11. *Campylobacter jejuni*	1,641,481	30.6	11	0	0
12. *Chlamydia trachomatis*	1,042,519	41.3	11	0	0
13. *Chlamydia muridarum*	1,069,412	40.3	11	0	0
14. *Chlamydophila pneumoniae* CWL029	1,230,230	40.6	11	0	0
15. *Chlamydophila pneumoniae* J138	1,228,267	40.6	11	0.01	0
16. *Chlamydophila pneumoniae* AR39	1,229,858	40.6	11	0	0
17. *Deinococcus radiodurans* R1 chromosome 1	2,648,638	67.0	11	0	0
18. *Deinococcus radiodurans* R1 chromosome 2	412,348	66.7	11	0.02	0
19. *Escherichia coli* K-12 MG1655	4,639,221	50.8	11	0	0
20. *Haemophilus influenzae* Rd	1,830,138	38.2	11	0	0
21. *Helicobacter pylori* 26695	1,667,867	38.9	11	−0.01	−0.01
22. *Helicobacter pylori* J99	1,643,831	39.2	11	0	−0.01
23. *Mycobacterium tuberculosis*	4,411,529	65.6	11	0	0
24. *Mycoplasma genitalium* G37	580,074	31.7	4	0.01	0
25. *Mycoplasma pneumoniae* M129	816,394	40.0	4	−0.02	0
26. *Neisseria meningitidis* MC58	2,272,351	51.5	11	0	0.01
27. *Neisseria meningitidis* Z2491	2,184,406	51.8	11	0	0
28. *Pseudomonas aeruginosa* PA01	6,264,403	66.6	11	0.01	−0.01
29. *Rickettsia prowazekii* Madrid E	1,111,523	29.0	11	0	0.01
30. *Synechocystis* PCC6803	3,573,470	47.7	11	0	0
31. *Thermotoga maritima*	1,860,725	46.3	11	0	0.02
32. *Treponema pallidum*	1,138,011	52.8	11	0	0.01
33. *Ureaplasma urealyticum*	751,719	25.5	4	0	0.02
34. *Vibrio cholerae* chromosome 1	2,961,149	47.7	11	−0.01	0
35. *Vibrio cholerae* chromosome 2	1,072,315	46.9	11	0	0.01
36. *Xylella fastidiosa*	2,679,306	52.7	11	−0.05	0.05

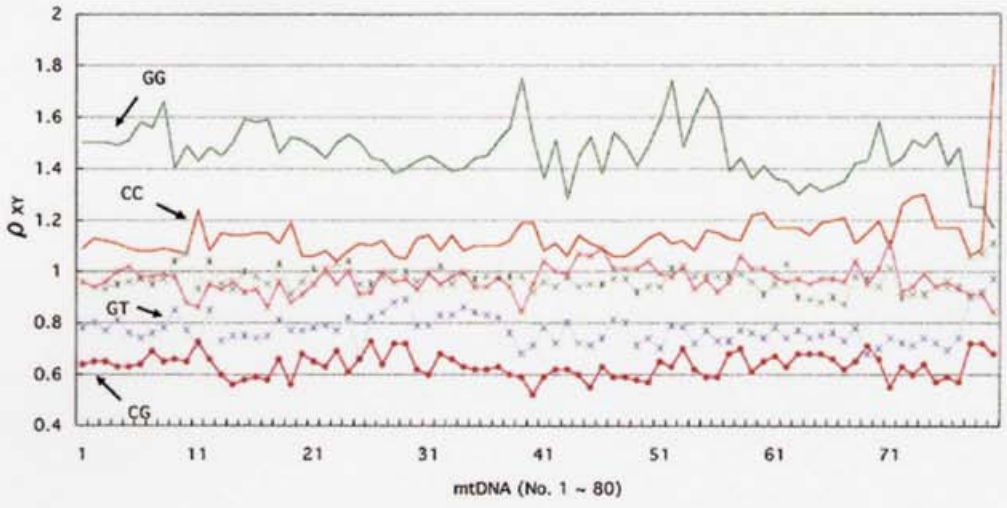
[a] The first 6 are archaebacteria and the remaining 30 are eubactaria. The total number of nucleotides over the 36 chromosomes is 70,046,804 bp.

ships are reversed. In other words, skews at the level of mononucleotides also manifest at the level of dinucleotides. Second, homo dinucleotide G_pG or C_pC is overrepresented in most mtDNAs, whereas both C_pG and G_pT are underrepresented, compared with the expected occurrences based on mononucleotide frequencies (Fig. 1). This trend has nothing to do with the presence of the methylation–deamination mechanism, since it is also found in organisms such as *Drosophila* which are believed not to possess the mechanism. The yeast mtDNA stands out again, showing enormous overrepresentation of G_pG and C_pC. Although the 86-kb genome is AT-rich and C_pG is underrepresented, each of G and C tends to occur in homo dinucleotides. Third, a deficiency of C_pG does not accompany an excess of T_pG or C_pA, which is expected to occur when the methylation–deamination process converts C to T in either strand. As noted by Cardon et al. (1994) and Karlin and Burge (1995), there must be some other mechanisms for generating these biases in dinucleotide frequencies.

Prokaryote Genomes

Interestingly, all 36 prokaryote chromosomes show that ATS and GCS in the entire region are almost exactly 0 (Table 2). This does not mean a complete absence of AT and GC skews. On the contrary, there exist strong skews in one half of the genome, but they are almost exactly canceled out by the opposite skews in the remaining half [Fig. 2; see also McLean et al. (1998) and references therein]. There is a more conspicuous rise and fall in GCS than in ATS over consecutive small windows, and the minimum and maximum of the cumulative GCS are located at the replication origin and terminus, respectively (Grigoriev 1998). The skews are thus local and attributed to the bidirectional mode of DNA replication.

Unlike mtDNAs, the frequency of a dinucleotide over the prokaryote genome is very similar to that of the inverted complementary dinucleotide (Burge et al. 1992). This symmetry holds true not only for hetero dinucleo-



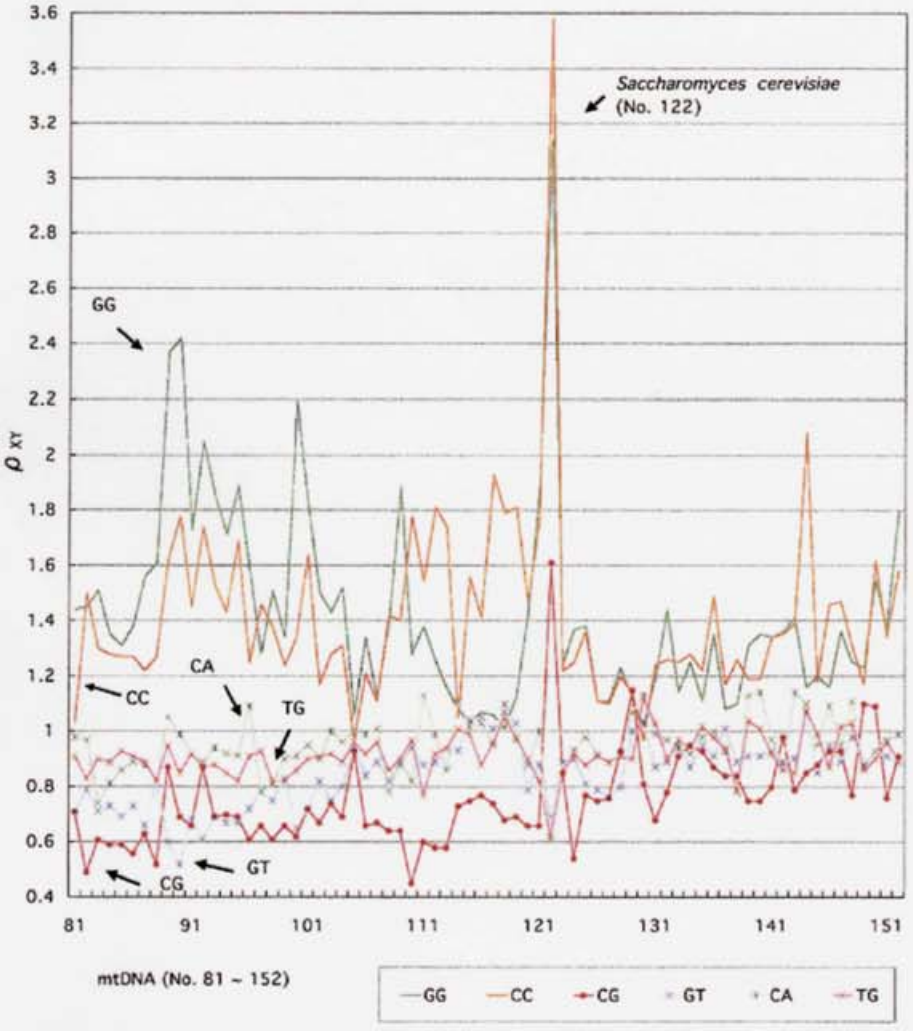


Fig. 1. Relative abundance [ρ_{XY} defined by (2)] of G_pG, C_pC, C_pG, G_pT, C_pA, and T_pG over the whole mitochondrial genome. The remaining dinucleotides showing more or less $\rho_{XY} = 1$ are suppressed. The 152 species are arranged along the abscissa in the same order as in Table 1. **Top:** 80 jawed vertebrates. **Bottom:** 72 invertebrates, fungi, plants, and protozoa. The unusual yeast mtDNA (No. 122) overrepresents not only G_pG and C_pC but also C_pG.

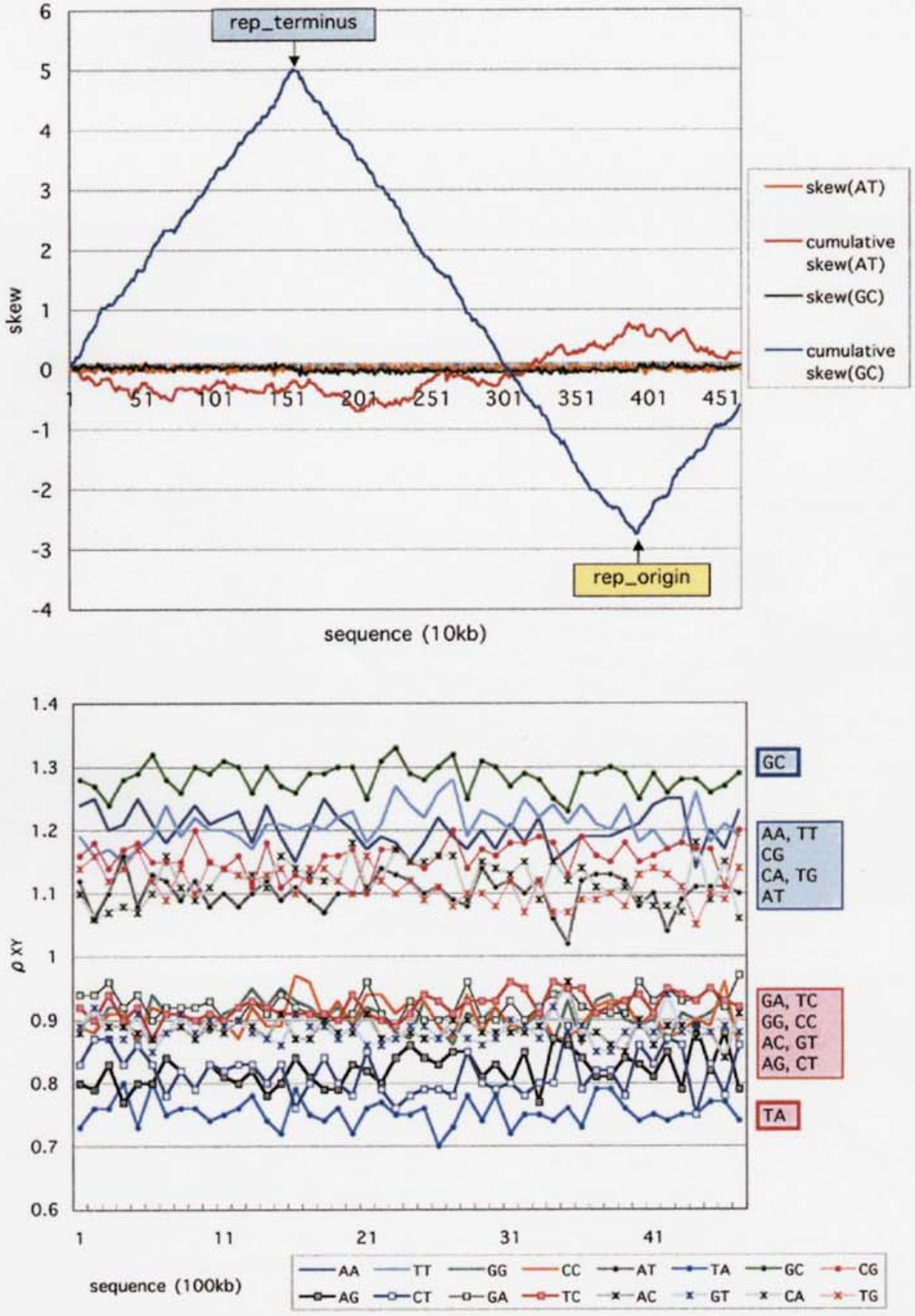


Fig. 2. AT and GC skews and relative abundance (ρ_{XY}) of 16 dinucleotides in the *E. coli* genome of 4.6 Mb. These are measured in windows (horizontal axis), each being either 10 kb (Top) or 100 kb long (bottom). **Top:** Cumulative values of ATS (*red*) and GCS (*blue*) over windows. The ATS and GCS values in individual windows become invisible compared with the cumulative values. The downward and upward *arrows* show the origin and terminus of DNA replication, respectively. **Bottom:** Relative abundance of 16 dinucleotides (ρ_{XY}). Note that C_pG is overrepresented, while T_pA is most underrepresented.

tides but also for homo dinucleotides, implying that any particular dinucleotide occurs equally frequently in both strands. To quantify skews of dinucleotide frequencies under strand asymmetry, it is convenient to define

$$\mathrm{DNS} = \frac{f_{XY} - f_{X'Y'}}{f_{XY} + f_{X'Y'}} \quad (3)$$

similar to (1). In the above, X′Y′ represents the inverted complementary dinucleotide of X_pY. The absolute value of DNS is smaller than 0.03 in all but one prokaryote genome. The exception is *Xylella fastidiosa,* belonging to the *Xanthomonas* group in the γ subdivision, and the genome is 20% more abundant in G_pT and T_pG than in their inverted complementary A_pC and C_pA, respectively. However, in general, no strand bias is found in the whole prokaryote genome, and this trend holds true even for tri- or tetranucleotides (data not shown).

Another rather surprising feature is that in some prokaryotes, the "universal" rule of C_pG deficiency is violated. That is, ρ_{CG} is greater than 1 and thus C_pG is overrepresented. The C_pG overrepresentation occurs in 13 eubacteria including *Bacillus subtilis, Escherichia coli,* and *Neisseria meningitidis,* but in none of the six archaebacteria. An example for the latter is the genome of *Aeropyrum pernix,* a member of desulfurococcaceae in Archaea, in which $\rho_{CG} = 0.70$, and an even more extreme example is the genome of *Methanococcus jannaschii,* in which $\rho_{CG} = 0.32$. In eubacterial genomes, the most underrepresented dinucleotide is T_pA or G_pT rather than C_pG, followed by A_pC and A_pT. It is interesting to note that the GC content (25.5%) of *Ureaplasma urealyticum* is the lowest among the 36 prokaryotes. Yet the ρ_{TA} value is 0.79, even smaller than $\rho_{CG} = 0.88$. Except for *A. pernix,* in which $\rho_{TA} = 1.21$, usage of T_pA is avoided in all other prokaryote genomes. These observations are inconsistent with the hypothesis of methylation–deamination–mutation (Karlin and Burge 1995).

Human Chromosomes 6p21.3, 21q, and 22q11.2-q13.2

There are many long contigs sequenced for the human genome. Here we use DNA sequences for a 4.4-Mb region of chromosome 6 containing HLA, a 28.5-Mb-long arm portion of chromosome 21, and a 23-Mb-long arm portion of chromosome 22. Of these, the HLA region is the best characterized concerning the gene density, C_pG island, functional analysis of genes, polymorphism, and so on.

Bernardi et al. (1985) and Ikemura (1985) found that chromosomal band zones are related to the mosaic structure of GC content and DNA replication timing. Subsequently, Ikemura and his co-workers have made detailed examinations of the HLA region, proposing that, among many, one GC junction boundary, between class III and class II gene clusters, is a chromosomal band boundary and corresponds to a switching point of DNA replication timing (Ikemura et al. 1990; Fukagawa et al. 1995). This junction is located 1 Mb away from the centromeric (right) end of the contig. Around this junction, ATS < 0 locally and the positive cumulative value decreases sharply, while GCS > 0 locally and the negative cumulative value approaches 0 (Fig. 3). In further telomeric regions, there are several prominent local maxma and minima in the cumulative ATS. Toward the telomeric end of class I region (left), there is also a reciprocal rise and fall between the cumulative GCS and the cumulative ATS values. By analogy, these local maxma and minima may correspond to replication origins.

The four homo dinucleotides occur more often than expected over the HLA region ($\rho_{XX} > 1$). It is interesting to note that f_{XX} reflects the GC content, but $\rho_{GG} \approx \rho_{CC}$ is greater than $\rho_{AA} \approx \rho_{TT}$ almost everywhere. Dinucleotide C_pG occurs with only 20–40% representation of the expected value, followed by T_pA, G_pT, A_pT, and A_pC. However, in small window sizes, a number of spikes of f_{CG} become visible. Most spikes correspond to unmethylated C_pG islands which are concentrated in the R-band chromosomal regions (Cross and Bird 1995).

Mono- and dinucleotide profiles of chromosomes 21 and 22 are presented in Figs. 4 and 5, respectively. The patterns of cumulative skews are different between the two chromosomes. In particular, the degree of skews is larger in chromosome 21 than in chromosome 22. Furthermore, in chromosome 21, the cumulative ATS and GCS are reciprocally ragged in the positive and negative region, respectively. The pattern is somewhat similar to that in mtDNAs. It is interesting to note that two-thirds of the chromosome 21 contig exhibits a low gene density and is extremely stable in dinucleotide frequencies. In chromosome 22, the cumulative ATS and GCS cross the horizontal axis several times, and this pattern is somewhat similar to that in the prokaryote genome.

In both chromosome 21 and chromosome 22, the 16 dinucleotides may be divided into four groups according to the value of ρ in Eq. (2). One group, having $\rho > 1$, is represented by A_pA, T_pT, G_pG, C_pC, T_pG, C_pA, A_pG, and C_pT. Like mtDNAs and prokaryote genomes, all four homo dinucleotides tend to be overrepresented. The second group, having $\rho \approx 0.8$, is represented by A_pT, T_pA, A_pC, and G_pT. The third group, with $\rho = 0.2$, is singly represented by C_pG. The final group, with $\rho \approx 1$, is represented by the remaining G_pA, T_pC, and G_pC. Like mtDNAs, a deficiency of C_pG is not compensated by an excess of T_pG, C_pA, or both, but unlike mtDNAs, the frequency of a dinucleotide is always equal or nearly equal to that of the inverted complementary dinucleotide (DNS = 0).

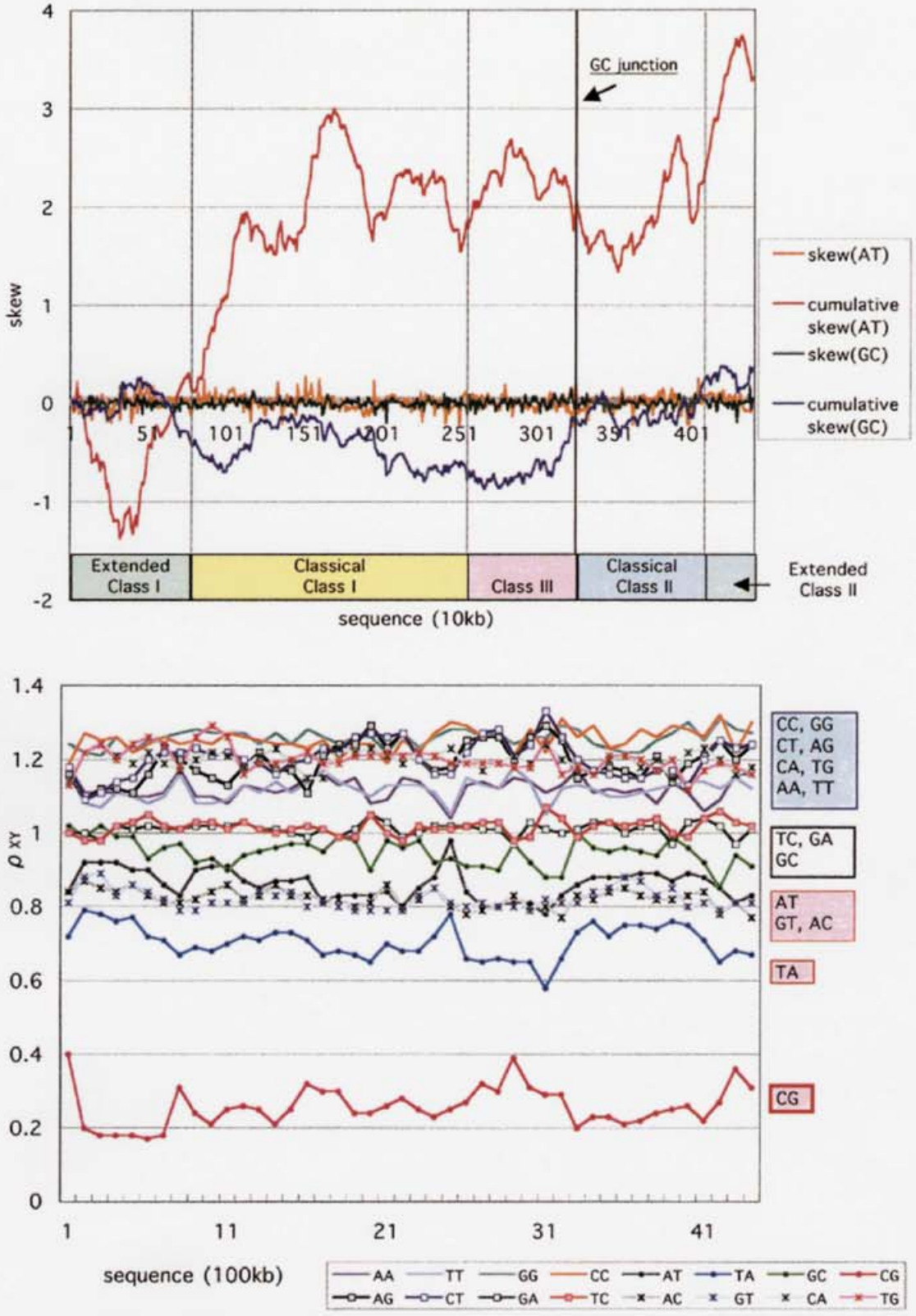


Fig. 3. AT and GC skews and relative abundance (ρ_{XY}) of 16 dinucleotides in the HLA region of 4.4 Mb. The **top** and **bottom** figures are comparable to those in Fig. 2, but the relative abundance (ρ_{XY}) of all 16 dinucleotides is presented. The GC junction corresponds to a switching point of DNA replication timing. In the classical class I, II (GC-poor), and III (GC-rich) regions, nucleotides A and C are more abundant than T and G, respectively, although dinucleotide A_pC and C_pG are underrepresented.

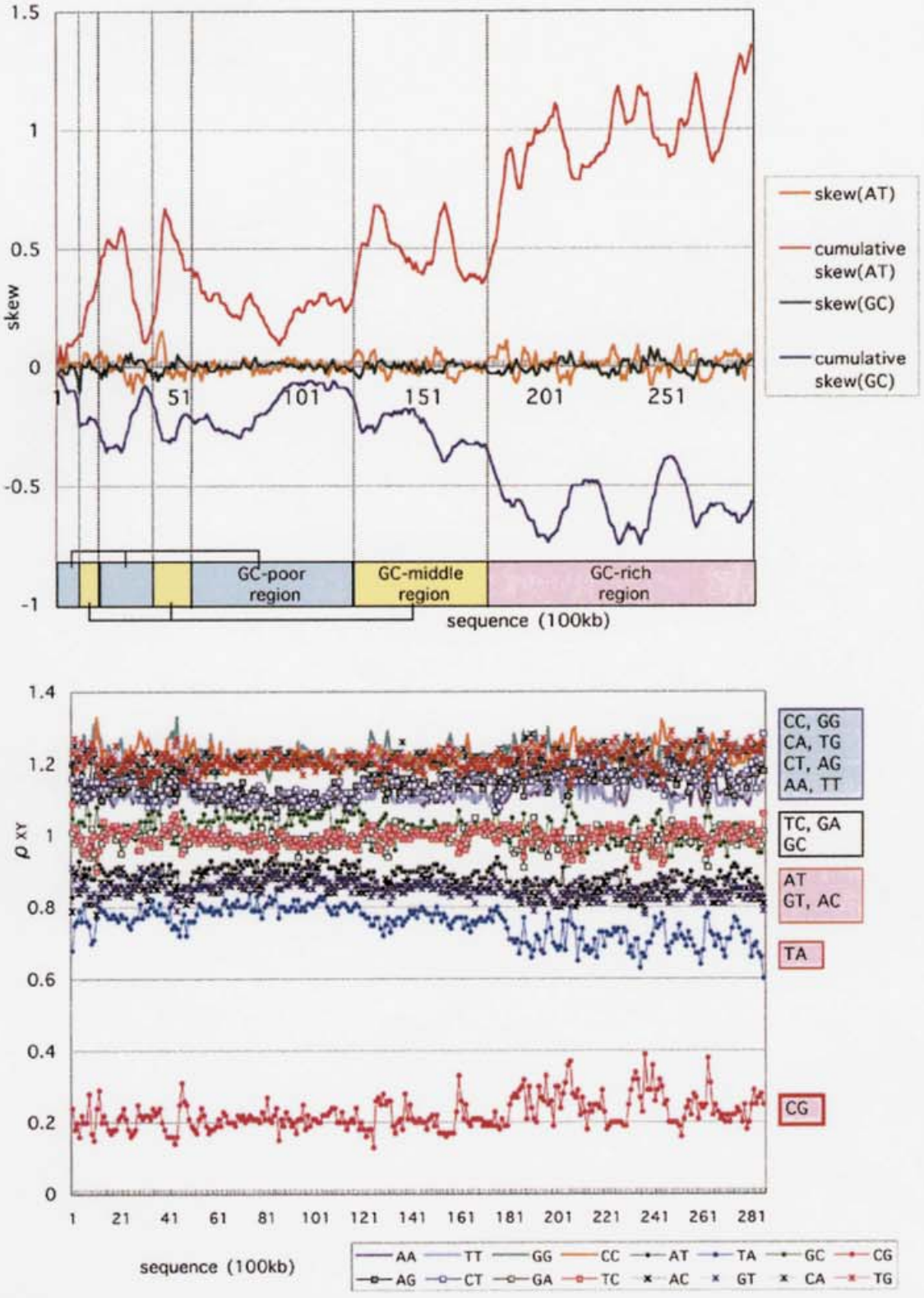


Fig. 4. AT and GC skews and relative abundance (ρ_{XY}) of dinucleotides in chromosome 21q of 28.5 Mb. The window size is 100 kb. The **top** and **bottom** figures are comparable to those in Fig. 3. Dinucleotides T_pA, A_pT, G_pT, and A_pC are more underrepresented and C_pG is less underrepresented in the GC-rich region than in the GC-poor region.

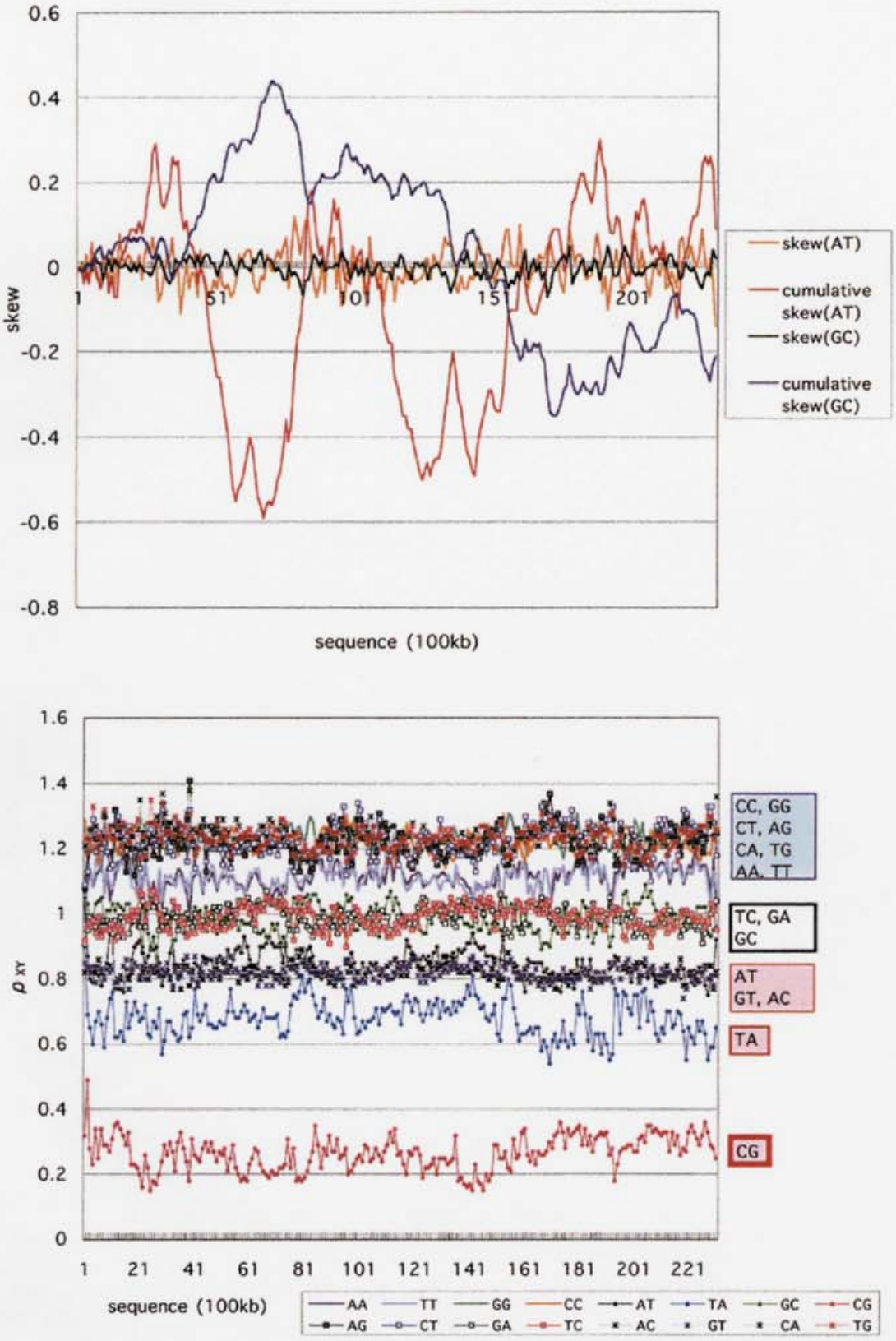


Fig. 5. AT and GC skews and relative abundance (ρ_{XY}) of dinucleotides in chromosome 22q of 23 Mb. The window size is 100 kb. The GC content varies greatly, ranging from 40 to 58%, but unlike chromosome 21 there is no GC-rich on poor region clearly clustered. The **top** and **bottom** figures are comparable to, but the scales of the ordinates are different from, those in Fig. 4.

Discussion

It is interesting to compare human chromosomes with other eukaryote chromosomes, so we examined a 17.5-Mb sequence of *A. thaliana* chromosome IV. Although the ATS and GCS profiles look like Fig. 4 for human chromosome 21, one remarkable difference is that C_pG is not very much underrepresented, and ρ_{CG} as well as ρ_{TA} is as high as 0.8. Another difference is the underrepresentation of G_pC. Such chromosome-specific patterns of mono- and dinucleotides cannot be easily explained, but the present analysis suggests that knowledge about strand asymmetry, genome structures, and DNA replication is indispensable. Indeed, all genomes show locally biased occurrences of mono- and dinucleotides which can be strongly influenced by strand-dependent replication errors. Francino and Ochman (1997) argued that both DNA replication and DNA transcription are asymmetric and can bias the occurrence of mutations between the complementary strands. Transcription may overexpose the nontranscribed strand to DNA damage. However, since the bias is found in noncoding regions as well, the highly asymmetric effect of transcription on mutagenesis alone cannot account for the characteristic dinucleotide profiles observed in eukaryote genomes.

The window size used here varies from one figure to another and is rather arbitrarily chosen. However, for eukaryote chromosomes, it must be properly adjusted in light of the size of a replicon. If the window size is too large, it is likely that strand asymmetry, even if exists, becomes undetectable. In particular, it is recalled that the window size used in analyzing human chromosomes is too large to demonstrate the C_pG island of about 1 or 2 kb often located in the 5′ region of genes (Cross and Bird 1995). If, on the other hand, the window size is too small, stochastic noise may erase all information. An appropriate window size also depends on whether di-, tri-, or polynucleotides are studied.

When we are interested in evolutionary changes or conservation of mono- and dinucleotide profiles, it is necessary to compare orthologous genomic regions among diverse organisms. As a first attempt, we compared four genetic loci, mostly among primates, but occasionally among mammals or even among eukaryotes: the β-globin gene cluster, BRCA1 (exon 11 only), Ikaros, and HSP70. The β-globin gene cluster includes introns and intergenic regions, while the remaining three are almost exclusively for exons. In terms of ρ_{CG}, Ikaros and HSP70 show rather high values in various species, in contrast to β-globin and BRCA1, with the ordinary low value of $\rho_{CG} = 0.2$. These exemplify that dinucleotide profiles are rather gene specific, although Ohno (1988) examined coding sequences then available and proposed a universal rule—T_pA/C_pG deficiency and T_pG/C_pT excess. Indeed, the deficiency is substantial for T_pA at the present four loci. However, the rule for C_pG is violated at Ikaros and HSP70, and the excess of T_pG/C_pT is slight or even absent at HSP70 in some species. Thus, the generality of Ohno's rule in coding regions has to be examined for various genes among diverse organisms. Our results for eukaryote chromosomes show that Ohno's rule, when applicable, is not restricted to coding regions, suggesting that the characteristic profile of dinucleotides stems from the intrinsic nature of DNA rather than from primordial repeating units in coding sequences.

Both single-strand DNA during replication or transcription and lagging-strand DNA during replication are more prone to errors than the complementary strand, as first noted by Wu and Maeda (1987) and Wu (1991). Since there is no a priori reason to assume that mutations take place in a similar manner between the two complementary strands, strand asymmetry should manifest not only in the rate but also in the pattern of nucleotide substitutions. In addition, the regional dependence of strand asymmetry can result in heterogeneous rates of nucleotide substitutions across the genome. These complications definitely violate some important assumptions made in the model of Jukes and Cantor (1969). Directly or indirectly, various refinements of their model have since been made in molecular evolutionary studies (e.g., Kimura 1980), but no doubt, Jukes and Cantor (1969) pioneered this then-new field of biology.

Acknowledgments. We thank Colm O'hUigin for his useful comments on an early version of this paper. This work was supported in part by Monbusho Grant 12304046 to N.T.

References

Beletskii A, Bhagwat AS (1996) Transcription-induced mutations: Increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli.* Proc Natl Acad Sci USA 93: 13919–13924

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Burge C, Campbell AM, Karlin S (1992) Over- and under-representation of short oligonucleotides in DNA. Proc Natl Acad Sci USA 82:1358–1362

Cardon LR, Burge C, Clayton DA, Karlin S (1994) Pervasive C_pG suppression in animal mitochondrial genomes. Proc Natl Acad Sci USA 91:3799–3803

Cross SH, Bird AP (1995) C_pG islands and genes. Curr Opin Genet Dev 5:309–314

Francino MP, Ochman H (1997) Strand asymmetries in DNA evolution. Trends Genet 13:240–245

Freeman JM, Plasterer TN, Smith TF, Mohr SC (1998) Patterns of genome organization in bacteria. Science 279:1827

Fukagawa T, Sugaya K, Matsumoto K, Okumura K, Ando A, Inoko H, Ikemura T (1995) A boundary of long-range G + C% mosaic domains in the human MHC locus: Pseudoautosomal boundary-like sequence exists near the boundary. Genomics 25:184–191

Furusawa M, Doi H (1992) Promotion of evolution: Disparity in the frequency of strand-specific misreading between the lagging and leading DNA strands enhances disproportionate accumulation of mutations. J Theor Biol 157:127–133

Furusawa M, Doi H (1998) Asymmetrical DNA replication promotes evolution: Disparity theory of evolution. Genetica 102/103:333–347

Grigoriev A (1998) Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res 26:2286–2290

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

Ikemura T, Wada K, Aota S (1990) Giant G+C% mosaic structures of the human genome found by arrangement of GenBank human DNA sequences according to genetic positions. Genomics 2:207–216

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism III. Academic Press, New York, pp 21–132

Jukes TH, Osawa S (1990) The genetic code in mitochondria and chloroplast. Experientia 46:1117–1126

Jukes TH, Osawa S (1993) Evolutionary changes in the genetic code. Comp Biochem Physiol 106B:489–494

Karlin S, Burge C (1995) Dinucleotide relative abundance extremes: A genome signature. Trends Genet 11:283–290

Karlin S, Mrázek J (1997) Compositional differences within and between eukaryotic genomes. Proc Natl Acad Sci USA 94:10227–10232

Kimura, M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kunkel TA (1992) Biological asymmetries and the fidelity of eukaryotic DNA replication. Bioessays 14:303–308

Lecrenier N, Foury F (2000) New features of mitochondrial DNA replication system in yeast and man. Gene 246:37–48

Lobry JR (1995) Properties of a general model of DNA evolution under no-strand-bias conditions. J Mol Evol 40:326–330

Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13:660–666

Lopez P, Forterre P, Guyader H, Philippe H (2000) Origin of replication of *Thermotoga maritima*. Trends Genet 16:59–60

McLean MJ, Wolfe KH, Devine KM (1998) Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J Mol Evol 47:691–696

Ohno S (1988) Universal rule for coding sequence construction: TA/CG deficiency–TG/CT excess. Proc Natl Acad Sci USA 85:9630–9634

Osawa S (1994) Evolution of the genetic code. Oxford Scientific, Tokyo

Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40:318–325

Takahata N, Kimura M (1981) A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. Genetics 98:641–657

Waga S, Stillman B (1994) Anatomy of a DNA replication fork revealed by reconstitution of SV40 DNA replication in vitro. Nature 369:207–212

Wu CI (1991) DNA strand asymmetry. Nature 352:114

Wu CI, Maeda N (1987) Inequality in mutation rates of the two strands of DNA. Nature 327:169–170