T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

題目(和文)	H-NSによる転写抑制を介した大腸菌ゲノムの多様性獲得機構
Title(English)	
著者(和文)	東光一
Author(English)	Koichi Higashi
出典(和文)	学位:博士（理学）, 学位授与機関:東京工業大学, 報告番号:甲第10087号, 授与年月日:2016年3月26日, 学位の種別:課程博士, 審査員:黒川 顕,山田 拓司,徳永 万喜洋,伊藤 武彦,山口 雄輝
Citation(English)	Degree:, Conferring organization: Tokyo Institute of Technology, Report number:甲第10087号, Conferred date:2016/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Powered by T2R2 (Science Tokyo Research Repository)

学位論文
平成27年度

H-NSによる転写抑制を介した大腸菌ゲノムの多様性獲得機構

東京工業大学
大学院生命理工学研究科
生命情報専攻
東　光一

指導教員　黒川　顕
山田　拓司

目 次

第 1 章　序論 **1**

1.1　本論文の背景 1

1.1.1　細菌の進化における遺伝子水平伝播の役割 1

1.1.2　核様体タンパク質 H-NS と遺伝子水平伝播との関係性 5

1.2　本論文の目的と構成 7

1.2.1　明らかにすべきこと 7

1.2.2　本論文の構成 10

第 2 章　H-NS 結合領域の保存性 **11**

2.1　緒言 11

2.2　材料と手法 13

2.2.1　大腸菌株 13

2.2.2　ChAP-seq 実験 17

2.2.3　シーケンシングリードのマッピングおよびデータの正規化 . 18

2.2.4　H-NS 結合領域の決定 20

2.2.5　大腸菌 3 株のゲノム上の“Common”領域の決定 22

2.3　結果 25

2.3.1　ChAP-seq 解析による大腸菌 3 株の H-NS 結合領域の比較 . . 25

2.3.2　“Common”領域上の H-NS 結合は株間できわめて精確に保存されている 27

2.3.3　H-NS 結合領域の局所的保存性 29

2.4　考察 31

第 3 章　H-NS 結合配列の多様化 **33**

3.1　緒言 33

3.2　材料と手法 36

3.2.1 大腸菌株 36
3.2.2 大腸菌オーソログ遺伝子の決定 38
3.2.3 オーソログ遺伝子の分子進化解析 41
3.2.4 H-NS 結合オーソログ遺伝子の定義の妥当性 41
3.2.5 Tree split compatibility による樹形比較 45
3.2.6 大腸菌オーソログ遺伝子のプロテオバクテリア門における保存性 48
3.2.7 大腸菌相同遺伝子間領域の決定 50
3.2.8 相同遺伝子間領域の分子進化解析 52
3.3 結果 53
3.3.1 H-NS 結合オーソログ遺伝子の多様化 53
3.3.2 進化解析の妥当性検証 1：H-NS 結合の定義 54
3.3.3 進化解析の妥当性検証 2：アラインメントの信頼性 55
3.3.4 進化解析の妥当性検証 3:H-NS 結合遺伝子の大腸菌種内 HGT の可能性 57
3.3.5 進化解析の妥当性検証 4：外来性遺伝子の特徴によるものか？ 59
3.3.6 H-NS 結合相同遺伝子間領域の多様化 66
3.4 考察 73

第 4 章 H-NS 結合配列の k-mer 特徴 78
4.1 緒言 78
4.2 手法 84
4.2.1 H-NS 結合ピーク位置の検出 84
4.2.2 ガンマ一般化線形モデルによるモデル化 86
4.3 結果 91
4.3.1 混合正規分布モデルによる H-NS 結合領域のモデル化およびピーク位置検出 91
4.3.2 ガンマ一般化線形モデルによる H-NS 結合強度のモデル化 . 94
4.4 考察 100

第 5 章 総括 103
5.1 本論文のまとめ 103

5.1.1　結果のまとめ 104
5.1.2　H-NS による転写抑制の進化的意義 105
5.2　今後の研究課題 109

参考文献 **112**

謝辞 **126**

図 目 次

2.1 大腸菌の進化系統樹 14
2.2 ChAP-seq 配列処理のフローチャート 19
2.3 実験間の H-NS 結合強度比較 21
2.4 H-NS 結合強度閾値の設定 23
2.5 Mauve による大腸菌 3 株の全ゲノムアラインメント 24
2.6 ChAP-seq 解析の再現性 26
2.7 株間 H-NS 結合領域の比較 28
2.8 H-NS 結合領域のオーバーラップ割合 29
2.9 H-NS 結合領域の局所的保存性 30

3.1 H-NS 結合遺伝子の発現量 35
3.2 Hidden paralog の概念 39
3.3 オーソログ候補遺伝子の大腸菌 44 株保存性 40
3.4 株特異的な配列の挿入による H-NS の結合 43
3.5 オーソログ遺伝子の多様性 53
3.6 H-NS 結合の定義を変えた場合のオーソログ遺伝子の多様性 56
3.7 ギャップなしアラインメントのみのオーソログ遺伝子の多様性 . . . 57
3.8 Minimal tree split compatibility 59
3.9 外来性オーソログとコアオーソログの多様性比較 60
3.10 プロテオバクテリア門における大腸菌オーソログの保存性 62
3.11 大腸菌オーソログ保存性の減衰 63
3.12 4 種類の大腸菌オーソログ遺伝子の発現量 64
3.13 相同遺伝子間領域の多様性 66
3.14 ギャップなしアラインメントのみの相同遺伝子間領域の多様性 . . . 67
3.15 相同遺伝子間領域の分類 68

3.16 class I および class II 相同遺伝子間領域の多様性 70
3.17 外来性遺伝子間領域とコア遺伝子間領域の多様性比較 71
3.18 プロモータ配列の有無による class II 遺伝子間領域の多様性比較 . . 72

4.1 A-tract 配列による DNA 二本鎖の屈曲 80
4.2 WS 配列によるドットマトリクス 83
4.3 DNA 配列と WS 配列によるドットマトリクスの比較 84
4.4 ガンマ分布確率密度関数のパラメータ依存性 90
4.5 混合正規分布による H-NS 結合領域のモデル化 92
4.6 染色体上の H-NS 結合強度の分布 93
4.7 ベイズ情報量基準によるモデル比較 95
4.8 モデルの予測値と観測値の比較 96
4.9 WS 5-mer モデルで学習された重みベクトルの係数 97
4.10 ATGC 3-mer モデルで学習された重みベクトルの係数 99

表 目 次

2.1 実験に使用したプライマー配列 16
2.2 各株の H-NS 結合領域のサイズ 27

3.1 分子進化解析に使用する大腸菌 44 株 37
3.2 株特異的な挿入・欠失に付随する H-NS 結合遺伝子 43
3.3 遺伝子保存性解析に使用した細菌株 49

4.1 本項で用いる記号 87

第1章　序論

本章では、遺伝子水平伝播という現象が大腸菌の進化に果たして来た役割を概観する。外来性遺伝子が宿主の遺伝子発現調節機構に組み込まれていく過程では、システムの安定性を維持するための何らかのメカニズムが必要であることに触れ、H-NS と呼ばれるタンパク質がその役割を担ってきたのではないかという、近年提案された仮説について説明する。本研究の目的はその仮説の実証である。最後に目的を達成するための要件をまとめ、本研究の構成を述べる。

1.1　本論文の背景

1.1.1　細菌の進化における遺伝子水平伝播の役割

細菌は、海底熱水噴出口 (Jannasch and Mottl 1985) から土壌 (Torsvik et al. 1990)、ヒト腸内 (Ley et al. 2006) にいたるまで、地球上のおよそありとあらゆる環境に生息し、多彩な代謝反応によって地球上の物質循環の根幹を担うとともに (Falkowski et al. 2008)、様々な高等生物を宿主とする共生関係によって宿主の健康や行動に密接に関わるなど (McFall-Ngai et al. 2013)、きわめて多様な代謝特性、細胞構造、および生理学的特徴を有した系統群である。このような細菌系統群の莫大な多様性は、それらのゲノムにおける点突然変異の蓄積のみによっては生み出され得ない。実際、ヒト腸内に共生する大腸菌と、ヒトに対して病原性を持つサルモネラ菌のゲノムを比較すると、それら共生菌と病原菌を区別するいかなる生理学的特徴も、両者の共通遺伝子に蓄積した突然変異には帰属させられない (Lawrence and Ochman 1998)。したがって、細菌が進化の過程で新たな環境に急速に適応し、その結果として細菌系統群の爆発的な多様化が実現された大きな要因は、遺伝子水平伝播（Horizontal gene transfer: HGT）と呼ばれる細菌系統群に特徴的な進化のメカニズムが、細菌の進化において支配的であったためだと考えられている (Ochman et al. 2000, Doolittle 2000)。

細菌の進化における HGT の重要性が最初に認識されたのは 20 世紀半ば、*Enterobacteriaceae* に属する細菌群において、ペニシリンへの耐性がプラスミドを介した HGT によって急速に広がっていることの報告に端を発する (Datta and Kontomichalou 1965)。その後、HGT は特定の薬剤耐性や病原性の獲得のみならず、細菌ゲノムの全域にわたって頻繁に生じていることが明らかとなった (Ochman et al. 2000)。細菌は、外界から積極的に遺伝子を取り込み、その遺伝子を自らのゲノムに組み込むことで、遺伝的多様性を獲得している。外来性遺伝子の水平伝播による遺伝的多様性の獲得は、細菌が環境条件の急激な変化に適応し、新たなニッチを獲得する手段として非常に重要である。

HGT によって獲得される遺伝子の多くは、ファージやプラスミド、トランスポゾンなどの可動性遺伝因子であるが、それら以外では特に、細胞表面構造に関わるタンパク質や、転写制御因子として機能する DNA 結合タンパク質、病原性に関わる遺伝子群が HGT によって伝播しやすいと考えられている (Nakamura et al. 2004)。HGT のもうひとつの重要な例は、CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) システムの伝播である。CRISPR は、リピート配列と、スペーサーと呼ばれるファージなどに由来する外来性配列が交互に並んだ構造を持った配列であり、それらのスペーサーと相補的な DNA 配列を有するファージの感染を防御する、細菌における獲得免疫システムである (Hatoum-Aslan and Marraffini 2014)。CRISPR は細菌系統群に広く存在するが、CRISPR 関連遺伝子群の進化系統関係が細菌系統群の進化と一致しないことや、コドン使用頻度のバイアス、GC 含量の違いなどから、CRISPR は多くの細菌において HGT によって獲得されたと考えられている (Horvath and Barrangou 2010)。CRISPR の獲得は、スペーサーに含まれる配列に関連したファージ群からの防御を一度に獲得することを意味するため、CRISPR 伝播の適応的意義はきわめて大きいと考えられる。

HGT によって、遺伝子を供与した細菌と遺伝子を受容した細菌が同じ遺伝子を共有することになるとはいえ、必ずしもそれらの細菌のニッチが重なることにつながるとは限らない (Cohan and Koeppel 2008)。例として、腸管毒素原性大腸菌（ETEC）のクラス 5 線毛が挙げられる。クラス 5 線毛は宿主細胞への接着に重要な定着因子の一種であり、ETEC はクラス 5 線毛を介して小腸上皮細胞上でコロニーを形成する。ETEC は、比較的近年の HGT によってクラス 5 線毛を獲得したことが知られ (Chattopadhyay et al. 2012)、同じくクラス 5 線毛を共有しているの

は、嚢胞性繊維症患者における肺疾患の原因として知られる *Burkholderia cepacia* であり、これらの種間での HGT が議論されている (Anantha et al. 2004)。すなわち、同じ定着因子を共有しながらも、ETEC は小腸上皮への定着にそれを利用し、*B.cepacia* は呼吸上皮への定着にそれを利用している。したがって、HGT は必ずしもニッチの収束を意味するのではなく、HGT によって獲得された遺伝子は、それを受容した細菌特有のニッチの文脈で評価され、新たなニッチの開拓に寄与したり、あるいは現在のニッチにおいてさらに優位性を拡大するために利用される (Cohan and Koeppel 2008, Wiedenbeck and Cohan 2011)。

HGT は、大腸菌の進化において特に重要な役割を果たしてきたと考えられている (Ahmed et al. 2008)。大腸菌は、ヒト腸内に共生するものから、病原性を示すものまで、非常に多様な生理学的特徴を有した多数の株によって構成される系統群である (Touchon et al. 2009)。大腸菌ゲノムの比較解析研究によって、それぞれの株のゲノムは、大腸菌株間で保存された共有ゲノム領域と、株特異的な領域によって構成されていることが明らかとなっている (Touchon et al. 2009)。病原性大腸菌の持つ遺伝子のうち、30%ほどは、比較的近年の HGT によって獲得されたものである (Koonin and Wolf 2008)。また、ヒト腸内共生菌の一種である大腸菌 SE11 株のゲノム解読の結果からは、実験室株の大腸菌 K-12 株のゲノムと比較して、およそ 1,000 個の遺伝子が SE11 株に固有であり、それらの HGT 遺伝子によって SE11 株は、腸管細胞上における定着の安定化という機能的優位性を得ていることが明らかとなった (Oshima et al. 2008)。このように、大腸菌は HGT によってゲノムの進化、種の多様化を果たし、新たなニッチを獲得してきたと考えられている。

HGT による遺伝子獲得には、以下の 3 つのステップが必要である (Ochman et al. 2000, Furuya and Lowy 2006)。第一に、接合、ファージの感染、あるいは DNA 分子の直接的な取り込みのいずれかによって、外来の DNA 分子を菌体内に取り入れる必要がある。第二に、取り入れた DNA 分子をゲノムに組み込む必要がある。第三に、ゲノムに組み込まれた外来性遺伝子の発現を適切にコントロールするための発現調節機構を確立する必要がある。三番目のステップは、特に重要である (Dorman 2007)。なぜなら、外来性遺伝子群がゲノムに取り込まれることで、これまでにない機能が獲得され、より環境に適した生存戦略を実現することができる一方で、外来性遺伝子の無作為な取り込みや、取り込まれた遺伝子の無作為

な発現は、高度に制御された生命システムの秩序を乱し、もはや生存を困難にする可能性もあるためである (Dorman 2007, Wiedenbeck and Cohan 2011)。HGT を考慮に入れた進化モデルの中には、遺伝子の取り込みと同時に即座に、増殖に有利な形質を細菌が獲得することを仮定したモデルがあるが (Niehus et al. 2015)、この仮定は多くの場合に適切ではない。細菌は外来性遺伝子の獲得以前に、それ自体として安定したゲノム構造と転写制御ネットワークを保持していたはずであり、既存のシステムの制御下にない“異物”としての外来性遺伝子の存在は、既存のシステムに擾乱を引き起こし、生存にとって大きなコストとなりうる。したがって、HGT によって外来性遺伝子を積極的に取り込み、多様性を増大させている細菌群は、取り込んだ遺伝子の発現をコントロールして生命システムの安定性を維持する何らかのメカニズムを有している必要がある。

獲得した外来性遺伝子の有害な効果を低減させるいくつかのメカニズムが考えられる。ひとつは補償的変異と呼ばれるもので、獲得した外来性遺伝子そのものに変異を導入して有害性を低減させるのではなく、その遺伝子と相互作用する既存の遺伝子が、外来性遺伝子の有害性に補償的に作用するように進化するプロセスである (Levin and Bergstrom 2000, Levin et al. 2000)。この現象は抗生物質耐性菌において観察されている。抗生物質がターゲットとする遺伝子は多くの場合転写や翻訳に関わるタンパク質であるが、抗生物質非存在下においてこれらの抗生物質耐性遺伝子はコストとなり、適応度を大きく下げてしまう。ストレプトマイシン耐性大腸菌のストレプトマイシン非存在下における進化実験では、抗生物質耐性であるリボソームタンパク質が抗生物質感受性のリボソームタンパク質に戻ってしまうのではなく、別のリボソームタンパク質に変異が導入されることによって、適応度低下を回避していることが明らかとなった (Levin et al. 2000)。

外来性遺伝子を既存のシステムに統合するためには、その有害性を低減させるだけでなく、転写量を適切に調整しなければならない。ひとつの方向性は、遺伝子重複などのメカニズムによって外来性遺伝子そのものの量を変化させることで、転写量を調整する方法である。遺伝子重複は新規に獲得された外来性遺伝子において特に生じやすく、それによって転写量を調整している可能性を示した研究がある (Hooper and Berg 2003)。一方で、細菌系統群の進化においては遺伝子重複は大きな役割は果たしてこなかったという報告もある (Treangen et al. 2011)。また、外来性遺伝子は多くの場合、既存の染色体上の塩基組成やコドン使用頻度とはか

け離れた配列的特徴を有しており、転写量の適切な調節のために、このような塩基組成やコドン使用頻度を既存のシステムに合わせて最適化するプロセス（“Amelioration”と呼ばれる）も観察されている (Lawrence and Ochman 1997)。いずれにしても、以上のような外来性遺伝子の最適化プロセスには進化のスケールで非常に長い時間が必要となる。その間、わずかにでも適応的なコストのかかる外来性遺伝子は、余計な遺伝子を排除する強力な選択圧によって、ゲノム上から喪失してしまう (Koskiniemi et al. 2012)。したがって、細菌が HGT によって多様性を獲得するためには、外来性遺伝子を単に獲得するだけではなく、獲得の初期段階において外来性遺伝子を特異的に認識して、その発現を抑制することによって有害な効果を低減し、またそれによって新規に獲得した遺伝子の喪失を防ぎながら、既存のシステムとの統合にいたる進化を生じさせるシステムの存在がきわめて重要となるのである。

1.1.2 核様体タンパク質 H-NS と遺伝子水平伝播との関係性

近年、大腸菌をはじめとした腸内細菌では、H-NS という DNA 結合タンパク質がこの役割の一端を担っていることが明らかとされつつある。H-NS は、大腸菌およびサルモネラ菌の持つ核様体タンパク質群の一種であり、核様体とよばれる細菌ゲノム高次構造の構成要素である。H-NS は、DNA 結合ドメインを持つと共に、ホモ多量体形成ドメインも持ち、細胞内では多量体を形成して (Ceschini et al. 2000, Smyth et al. 2000, Kotlajich et al. 2015)、特定の遺伝子のプロモーター領域に結合することによって (Spurio et al. 1997, Ueguchi et al. 1997, Rimsky 2004, Dame et al. 2005)、それらの遺伝子の発現を抑制していると考えられている。これまで、H-NS のゲノム上局在化領域の解析によって、H-NS は大腸菌やサルモネラ菌のゲノム上の数百か所に結合することが明らかにされてきた (Grainger et al. 2006, Lucchini et al. 2006, Navarre et al. 2006, Oshima et al. 2006, Kahramanoglou et al. 2011)。H-NS 結合配列の比較から、H-NS は A(Adenine) および T(Thymine) を多く含む DNA 配列（AT-rich 配列）に結合しやすい傾向があることが見出されている (Grainger et al. 2006, Lucchini et al. 2006, Navarre et al. 2006, Oshima et al. 2006, Kahramanoglou et al. 2011)。さらに、AT-rich 配列中の TA、たとえば ATATA といった配列が、H-NS に認識されやすい、高親和性の DNA 配列だとい

うことが明らかにされている (Lang et al. 2007, Sette et al. 2009, Kahramanoglou et al. 2011, Gordon et al. 2011)。また、H-NS はこれらの高親和性配列を最初に認識して結合し、そこをコアとして H-NS 多量体を形成して周囲に広がっていくという、「核形成・伝播」モデルが提案されている (Lang et al. 2007)。

H-NS は、AT-rich な DNA 配列に結合し、遺伝子の発現を抑制するという機能を持つ。一方、大腸菌やサルモネラ菌のゲノム配列全体の AT 含量はおよそ 50%であり、これらの細菌にとっての外来性遺伝子は、AT-rich であることが多い。そのため H-NS は、HGT によって新規に獲得された遺伝子の発現によって生じる問題を回避するための重要な因子であると考えられている (Lucchini et al. 2006, Navarre et al. 2006, Oshima et al. 2006, Dorman 2007, Navarre et al. 2007)。

大腸菌の進化過程における H-NS の役割として、以下のようなモデルが提案されている。外来性遺伝子が大腸菌ゲノムに組み込まれた時、それらが含む AT-rich な DNA 配列を H-NS が認識し、多量体を形成することで外来性遺伝子領域を覆ってしまい、それらの遺伝子の発現を抑制する。それによって、既存のシステムを撹乱することなく、外来性遺伝子が安全にゲノム内に保持される。その後、環境条件の変化によって外来性遺伝子の発現が必要となったときには、H-NS の結合と競合する転写調節因子によって、それらの遺伝子の発現が活性化される。したがって、H-NS が外来性遺伝子の獲得時に発現を抑制してしまうことの意義は、それらの遺伝子を安全に保持しながら、いずれ適切に“使いこなす”ことができるように、転写調節システムを進化させるための時間的猶予を与えることにある。H-NS による発現抑制から遺伝子を“解き放つ”ための転写調節システムを構築するためには、次のいずれかの方法を取らなければならない (Dorman 2007)。

1. 外来性遺伝子のプロモーター領域に、たまたま宿主の持つ転写活性化因子の結合領域が存在している。

2. 宿主の持つ転写活性化因子が、外来性遺伝子の発現を活性化させられるように適切に進化する。

3. 外来性遺伝子のプロモーター領域が、宿主の持つ転写活性化因子によって認識されるように進化する。

一つめの可能性、すなわち外来性遺伝子がゲノムに組み込まれた段階ですでに、その遺伝子を適切に制御できるシステムが整っているという可能性は、獲得した

遺伝子が系統的に非常に近縁な細菌由来のものでない限り、きわめて低い。さらに、適切な制御のためには、獲得遺伝子上に転写活性化因子の結合領域が存在しているだけでは不十分であり、転写調節因子による活性化が H-NS による転写抑制と競合するような、転写調節因子間の適切な関係性が構築されなければならない。また、二つめの進化経路は、細菌のようにコンパクトなゲノムを持った生物種にとってはコストの大きい戦略である。ゲノム上の数百個の外来性遺伝子はそれぞれ独自の配列を有しており、個別に最適化された転写調節因子を生み出していく場合、獲得遺伝子の数の増加に応じて必要とされる転写調節因子の数が増大していってしまう。最節約的に進化を考えた場合、最も可能性として高いのは三つめの進化経路である。なぜなら、この進化戦略であれば、比較的狭い領域の配列の変異によって、既存の転写調節ネットワークと新規獲得遺伝子との統合が達成され得るためである。実際、大腸菌とサルモネラ菌の系統において、外来性遺伝子の網羅的な分子進化解析を行った研究では、新規に獲得された外来性遺伝子は、大腸菌およびサルモネラ菌のコア遺伝子と比較した場合に、より少数の転写調節因子に制御されており、さらに、外来性遺伝子の獲得年代が古くなるにしたがって制御される転写調節因子の数が増大していることが示唆されている (Lercher and Pál 2008)。この結果は、外来性遺伝子がゲノムに組み込まれ、ゲノムに“馴染んで”いく過程で、転写調節領域の微調整が行われ、既存の転写調節ネットワークとのリンクが増大していくことを意味している。本論文で、実際の大腸菌ゲノムデータから実証を試みるのは、この三つめの進化仮説、すなわち、このような進化過程に H-NS が積極的役割を果たしているという仮説である。

H-NS はこのように、外来性遺伝子の転写調節システムを適切に構築するための時間的猶予を宿主に与える役割があると考えられている。H-NS による外来性遺伝子の発現抑制機能によって大腸菌は、外来性遺伝子を獲得して遺伝子プールの多様性を保持しつつ、様々な環境条件に適応する機会を得ている可能性がある。

1.2 本論文の目的と構成

1.2.1 明らかにすべきこと

H-NS は、大腸菌が進化し、多様性を獲得していく過程に深く関わってきたと考えられる。分子進化解析に基づいた研究により、外来性遺伝子が宿主の転写調節

ネットワークに統合される過程において、転写調節領域に変異を蓄積させることによって、その発現様式が微調整されてきたことが示唆されている (Lercher and Pál 2008)。また、これまでに大腸菌、サルモネラ菌などいくつかの種において、Chip-chip 解析および Chip-seq 解析によって、H-NS の局在化領域が明らかにされてきた (Grainger et al. 2006, Oshima et al. 2006, Kahramanoglou et al. 2011)。これらの研究では、大腸菌ゲノムにおいて、外来性遺伝子が組み込まれた領域に H-NS が結合しやすい傾向があることを指摘しているが、外来性遺伝子が組み込まれて“以降”のその領域の進化過程が、コア遺伝子の進化過程とどのように異なるのか、そしてそれが H-NS の結合とどのように関係しているのかについては検討していない。遺伝子水平伝播は、けっして外来性遺伝子の染色体への組み込みのみによって完結する現象ではない。既存の転写調節ネットワークとの統合および発現様式の微調整が必須であること (Dorman 2007, Wiedenbeck and Cohan 2011) を考えると、細菌の爆発的多様化を支えてきた遺伝子水平伝播という現象を包括的に理解するためには、大腸菌系統内の詳細な分子進化解析と、ゲノム上の H-NS 結合領域の解析とを結びつける研究が必須である。本論文では以下に挙げる 3 つの問題点について、比較ゲノム解析、データ可視化、分子進化解析、多変量解析などの情報解析手法を用いて取り組み、H-NS の本質的な役割の理解を目指す。

H-NS の結合は進化の過程で保存されるのか？

大腸菌のゲノムは、遺伝子水平伝播を介してダイナミックかつ可塑的な進化を遂げてきたが (Hayashi et al. 2001)、その過程でゲノム上の H-NS の結合がどのように変化してきたのか、あるいは維持されてきたのかは、明らかになっていない。これまでの研究のような単一の種の H-NS 結合位置情報だけから、ゲノムと核様体構造の協調的な進化・多様化の過程を類推する事は困難である。H-NS がゲノムの多様性創出にどのように関わってきたのかを理解するためには、複数種間のゲノム上 H-NS 結合位置の保存性・多様性について、ゲノム構造の進化的な観点に立脚した比較解析を行う必要がある。そこで本研究では、新型シーケンサーを用いて、異なった系統群に属する 3 株の大腸菌、実験室株の K-12 株 (Hayashi et al. 2006)、ヒト腸内共生細菌の SE11 株 (Oshima et al. 2008)、同じくヒト腸内共生細菌だが K-12 株や SE11 株とは系統的に遠い SE15 株 (Toh et al. 2010) のゲノム上の H-NS 結合領域を決定し、それらの分布を比較する。3 株のゲノム上の H-NS 結

合領域を比較するための情報解析パイプラインを開発し、H-NS 結合領域の進化的保存性・多様性を調べる。

H-NS 結合領域下の DNA 配列はどのような進化的特徴を持っているのか？

H-NS の結合が大腸菌ゲノムの多様化に影響を及ぼしてきたのであれば、H-NS 結合領域下の DNA 配列と、H-NS が結合していない DNA 配列それぞれで分子進化解析を実行し、その結果を比較することで、H-NS 結合 DNA 配列に特異的な何らかの進化的特徴が観察できるはずである。前節で述べた H-NS の役割に関する Dorman の仮説 (Dorman 2007) が真実であるならば、H-NS 結合 DNA 配列では、それぞれの大腸菌の生息環境条件に応じた最適化がなされており、したがって、配列多様性が増大していることが予想される。また、その進化の特徴を詳細に分析することによって、配列多様性が H-NS 結合による積極的な変異の導入によってもたらされたのか、あるいは H-NS による転写抑制の間接的な影響によって生じたものなのかを区別することができる。しかし、大腸菌はゲノム構造として多様であるとはいえ、進化系統的には非常に近縁な種から構成される系統群であるため、先述の 3 株のみの分子進化解析から得られる結果では、H-NS が進化に及ぼす影響を観察するにあたって解像度として不十分である。そこで本研究では、公共データベースに登録されている 44 株の大腸菌ゲノム情報を収集し、それらについて徹底的な比較ゲノム解析を実施することによって、高解像度の分子進化解析および H-NS 結合領域分布との比較解析を実行する。

H-NS 結合領域下の DNA 配列はどのような配列的特徴を持っているのか？

H-NS が外来性遺伝子に結合し、変異が蓄積して、その結果としてただちに H-NS の結合が失われたとしたら、Dorman の仮説 (Dorman 2007) が提唱するような、外来性遺伝子が宿主の転写調節ネットワークに組み込まれるまでの時間的猶予を稼ぐという H-NS の役割は実現できない。H-NS がそのような役割を果たすためには、H-NS が配列上の厳密な結合モチーフに基づいて染色体に結合するのではなく、ある程度の変異を許容するような、柔軟な結合モチーフに基づいて結合している必要がある。さらに、H-NS 結合の「核形成・伝播」モデル (Lang et al. 2007) にみられるように H-NS の結合はダイナミックであること、および、競合する転

写活性化因子 (Dorman 2007, Fang and Rimsky 2008) あるいは浸透圧や酸性ストレスなどの環境シグナルによって H-NS の結合が解かれ得る (Atlung and Ingmer 1997, Hommais et al. 2001, Dorman 2004) ことを考えると、H-NS のゲノム上結合“強度”は、その配列上の H-NS 結合モチーフの存在、あるいは結合モチーフの変化の度合いを反映していると考えることができる。本研究では、H-NS の結合強度パターンを混合ガウス分布モデルで近似することによって、H-NS 結合の“核”となりうる領域を抽出し、さらに、それらの領域に含まれる配列の k-mer 特徴量のデータから H-NS 結合強度を予測するための回帰モデルを構築することによって、H-NS の結合に本質的に寄与していると推定される配列的特徴の特定を試みる。

1.2.2 本論文の構成

本論文の構成は次の通りである。第 2 章では、本論文のすべての基礎となる ChAP-seq 解析の手法について説明し、H-NS 結合領域情報を取得するために開発した解析パイプラインおよび比較 ChAP-seq 解析手法の詳細について述べる。また、その結果得られたデータから、H-NS 結合領域の進化的な保存性・多様性を議論する。第 3 章では、進化解析の対象とする大腸菌株を 44 株に増やし、高精度の分子進化解析を行うことによって、H-NS 結合遺伝子および H-NS 結合遺伝子間領域において、特徴的な進化が生じていることを見る。その結果から、H-NS の役割として提唱されている仮説が観察される現象をよく説明していることを議論する。第 4 章では、H-NS 結合領域の配列的特徴について詳細に検討する。H-NS 結合配列に見られる特異的な配列パターンの発見を試みる。第 5 章では、以上の結果を総合し、今後の展望について述べる。

第2章　H-NS結合領域の保存性

本章では、異なる大腸菌株間のH-NS結合の保存性について検討する。H-NS結合領域を高精度に決定するChIP-seq法について詳述し、その結果得られたデータを比較ゲノム解析と組み合わせることによって、異なるゲノム間のDNA結合領域情報を比較するための比較ChIP-seq解析の手法を述べる。データ可視化、統計情報抽出の結果から、大腸菌の進化過程におけるH-NS結合の保存性を議論する。

2.1　緒言

これまで、大腸菌やサルモネラ菌などいくつかの種について、ChIP-chip法やChIP-seq法などの、DNA結合タンパク質のゲノム上結合位置を高精度に決定する実験手法に基づき、ゲノムレベルのH-NS結合分布が調べられてきた (Grainger et al. 2006, Oshima et al. 2006, Kahramanoglou et al. 2011)。それらの研究結果からH-NSの結合領域は、染色体上でも特に、水平伝播によって獲得されたと推定される外来性遺伝子がコードされている領域周辺に偏って存在していることが明らかにされてきた。しかし、異なる菌株間での比較解析に基づいて、ゲノム上のH-NS結合領域の保存性・多様性を検討する研究はこれまで行われていない。大腸菌はとりわけ、ゲノム構造の可塑性、ゲノム配列の多様性が高い系統群である。染色体サイズだけを考えても、たとえば大腸菌O157株とK-12株を比較した場合、O157株は25%ほど染色体サイズが大きい (Hayashi et al. 2001)。一方で、染色体の基本骨格と考えられる領域は配列としてきわめて安定であることが知られており、菌株特異的配列はこの基本骨格に様々なサイズで挿入された外来性DNA配列であると考えられている (Hayashi et al. 2001, Ogura et al. 2009)。したがって、H-NSの結合が大腸菌ゲノムの進化の過程でどのように変化してきたのか、あるいは維持されてきたのかを明らかにするためには、まず大腸菌複数株間の比較ゲノム解析を行い、その結果得られたゲノム構造自体の保存性・多様性の情報をベー

スとして、それぞれの株の H-NS 結合領域を比較しなければならない。

本研究では、比較対象の大腸菌株として、実験室株の K-12 株（A 群大腸菌）(Hayashi et al. 2006)、ヒト腸内共生細菌の SE11 株（B1 群大腸菌）(Oshima et al. 2008)、同じくヒト腸内共生細菌だが K-12 株や SE11 株とは系統的に遠い SE15 株（B2 群大腸菌）(Toh et al. 2010) を用いる。本章で決定するそれぞれの株の H-NS 結合領域の情報は、本研究全体のベースとなる重要なデータであり、可能な限り実験ノイズの影響が除去された、精度の高い結合領域情報を取得しなければならない。本研究では、H-NS 結合領域を特定する手法として、DNA 結合タンパク質と共沈降する DNA 断片を新型シーケンサーで網羅的に読み取り、ゲノム上の広範囲に及ぶ結合領域の特徴を決定することのできる ChIP-seq 法を用いた。また H-NS は、ゲノム上の特定の位置に結合する際、単一のタンパク質として結合するのではなく、H-NS 多量体を形成して結合し、高次の H-NS-DNA 複合体を形成する。そのため、ChIP-seq のデータを解析することによって得られる H-NS の結合位置は、単に結合強度波形のピーク位置を求めるのではなく、H-NS 結合のゲノム上の広がりを勘案した“H-NS 結合領域”として決定する必要がある。H-NS はゲノム上の数百か所に結合し (Grainger et al. 2006, Oshima et al. 2006, Kahramanoglou et al. 2011)、それぞれの H-NS 結合領域の長さは、数百塩基から数万塩基にいたるまで、大きく異なる。そこで本研究ではまず、ChIP-seq データを解析して、H-NS が結合していると考えられるゲノム上の範囲を取得する手法を開発し、統計的に有意な H-NS 結合領域を高精度に同定する手法を開発した。

前述のように、3 株の大腸菌は異なったゲノム構造を持つことから、それらのゲノム上の H-NS 結合領域を直接に比較することはできない。そこで、H-NS 結合領域の比較に先立って、3 株の大腸菌が共有する“Common”ゲノム領域と、各株のゲノムに固有の“Specific”ゲノム領域を特定することを目的として、まず 3 株のゲノムについて、全ゲノムアラインメントを行った。シンテニーの保存を加味して、“Common”ゲノム領域を慎重に特定したことで、このゲノム領域は、少なくとも 3 株の大腸菌が種分岐する直前には確立されていたゲノム領域だと考えられる。また、“Specific”領域は、3 株でそれぞれ、比較的近年に獲得された外来性因子によって構成されたゲノム領域であると考えられる。“Common” ゲノム領域と、それ以外の、3 株中 2 株で共有されている“Shared”ゲノム領域、および、各株が固有に保持している“Specific”ゲノム領域上の H-NS 結合領域の分布を比較

することによって、先行研究が示唆しているように、H-NS の結合は近年獲得された外来性遺伝子領域に限られるのか、あるいは、H-NS の結合が長期的に維持される場合があるのかを検討した。

本章の構成は次の通りである。2.2 節では、本研究で対象とした大腸菌株の情報、ChIP-seq 法、ChIP-seq データの情報解析による結合領域の特定方法、比較ゲノム解析の手法について述べる。2.3 節では、これらの手法に基づいて行われた比較 ChIP-seq 解析の結果をまとめる。ここでは特に、“ Common ”、“ Shared ”、“ Specific ” それぞれの染色体領域にみられる H-NS 結合パターンについて詳細に検討する。2.4 節では、結合パターンの比較結果から、大腸菌の進化過程における H-NS 結合の保存性について議論する。

2.2　材料と手法

以下、実験操作（第 2.2.1 節から第 2.2.2 節）に関しては共同研究者が実施した。情報解析（第 2.2.3 節から第 2.2.5 節）は筆者が単独で行った。

2.2.1　大腸菌株

本研究で用いた大腸菌株は、実験室株の K-12 株（A 群大腸菌）、ヒト腸内共生細菌の SE11 株（B1 群大腸菌）、同じくヒト腸内共生細菌だが K-12 株や SE11 株とは系統的に遠い SE15 株（B2 群大腸菌）の 3 株である（図 2.1）。3 株はいずれも大腸菌の一種であるが、図 2.1 からわかるように、SE15 株は進化系統的に SE11 株、K-12 株とは離れた株であり、それらの間の進化距離は、現時点で大腸菌系統群に属することがわかっている細菌種間の進化距離の中でも最大に近い。したがって、SE15 株と他の 2 株のゲノム情報、H-NS 結合情報を比較することにより、大腸菌系統群全体で生じている多様性をおおよそ見積もることが可能だと考えられる。また SE11 株は、K-12 株と進化的に近縁であるが、約 1,000 個の遺伝子を固有に持ち、腸管細胞上における定着の安定化など、K-12 株とは異なった生理学的特徴を有していることが知られている (Oshima et al. 2008)。この 3 株を比較対象とすることで、大腸菌という種の成立時点から現在にいたるまで、幅広い進化スケールにおける H-NS 結合の安定性を議論することが可能となる。なお、3 株はいずれも *hns* 遺伝子を保持し、H-NS タンパク質のアミノ酸配列は完全に一致している。

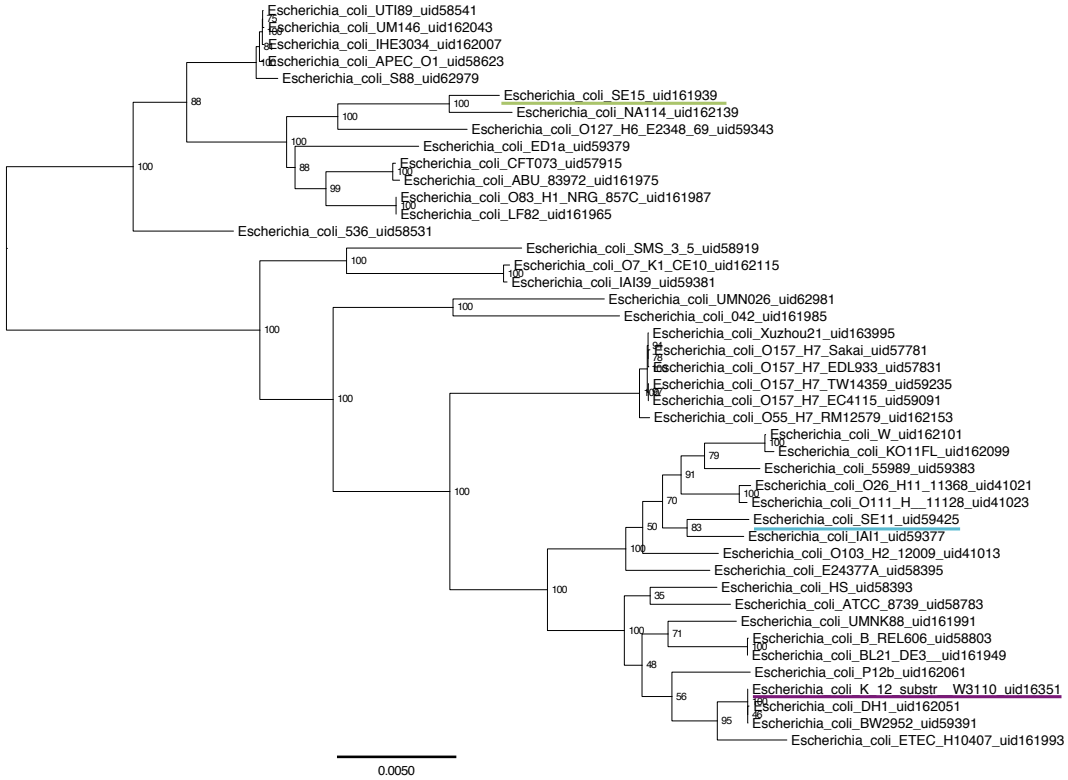

図 2.1: 公共データベースに登録されている大腸菌 44 株のゲノム情報を用いて作成した進化系統樹。緑色の下線が SE15 株、水色の下線が SE11 株、紫色の下線が K-12 株を示す。系統樹は第 3 章で述べる手順によって推定した。

本研究では、H-NS のゲノム上結合位置を高精度に特定するために、ChIP 法（クロマチン免疫沈降法）を改良した ChAP 法（クロマチンアフィニティ沈降法）を行った (Ishikawa et al. 2007, Park 2009)。これは、DNA 結合タンパク質と共沈降する DNA 断片を、DNA 結合タンパク質の抗体で捕捉するのではなく、His タグを付加した DNA 結合タンパク質を発現させることによって、DNA 断片の回収効率を大きく高めた手法である。C 末端に His12 タグ (12 x Histidine) を付加した H-NS を発現する大腸菌は、以下の方法で構築した。

K-12(W3110)H-NS-His12 株を構築するために、まず、His12 タグを 3’ 末端に付加した *hns* 遺伝子の DNA 断片を、one-step gene inactivation method (Datsenko and Wanner 2000) を改良した手法を用いて大腸菌 BW25113 株に導入した。His12 タグとカナマイシン耐性遺伝子を含んだ配列は、TOP705 および TOP706 プライマー（表 2.1）を使用して、pSTV28-C-12His をテンプレートとした PCR によって増幅した。pSTV28-C-12His は、pKD4 由来のプラスミドであり、12 個の連続した Histidine 配列と、それに隣接したカナマイシン耐性遺伝子を持っている (Datsenko and Wanner 2000)。BW25113 H-NS-His12 株の構築後、P1 トランスダクションによって H-NS-His12 を K-12(W3110) 株に導入した。

H-NS-His12 の SE11 株、SE15 株への導入は、pDEX プラスミドを用いた gene doctoring method (Lee et al. 2009) を利用した。pSTV28-C-12His 上の H-NS-His12 とカナマイシン耐性遺伝子を含んだ DNA 断片を増幅するために、SE11 株においては hns-His12-H1 と hns-His12-H2-1 プライマーを、SE15 株においては hns-His12-H1 と hns-His12-H2-2 プライマーを用い（表 2.1）、GXL DNA polymerase(Takara Bio, Japan) を用いて PCR を行った。増幅した DNA 断片は pDEX プラスミドの EcoRV site に挿入した。SE11 株、SE15 株をそれぞれ、pACBSR と pDEX の二つのプラスミドで同時形質転換した。形質転換体は、0.2%アラビノースと 25μg/ml クロラムフェニコールを含んだ LB 培地で数時間培養した。その後、5%スクロースを含んだ LB 培地、30 ℃でさらに二時間培養し、50μg/ml カナマイシンと 5%スクロースを含んだ LB プレートで、カナマイシン耐性かつスクロース耐性のコロニーを選択した。すべての形質転換体について、H-NS-His12 の発現は、抗 His タグ抗体 (MBL, JAPAN) を用いたウェスタンブロッティングによって確認した。また、すべての形質転換体について、H-NS-His12 の導入が好気条件下の LB 培地における細胞増殖に影響を与えないことを確認した。

表 2.1: 実験に使用したプライマー配列

ID	sequence	comment
hns-His12-H1	AAGCAATGGATGAGCAAGGTAAATCCCTCGACGATTTCCTGATCAAGCAAAGAGGATCGCATCACCATCA	construction of C-terminal His12 fused H-NS in SE11 & SE15
hns-His12-H2-1	GTCTTAAACCGGACAATAAAAAATCCCGCCGCTGGCGGGATTTTAAGCAACATATGAATATCCTCCTCCTTAG	construction of C-terminal His12 fused H-NS in SE11
hns-His12-H2-2	GTCTTAAACCGGACAATAAAAAATCCCGCCGATGGCGGGATTTTTAGCATCATATGAATATCCTCCTCCTTAG	construction of C-terminal His12 fused H-NS in SE15
TOP705	ACTCCAGCTGTAATCAAAAAAGCAATGGATGAGCAAGGTAAATCCCTCGACGATTTCCTGATCAAGCAAAGAGGATCGCATCACCATCACCATC	construction of C-terminal His12 fused H-NS in K-12(W3110)
TOP706	ATTCTTATTAAATTGTCTTAAACCGGACAATAAAAAATCCCGCCGCTGGCGGGATTTTAAGCAACATATGAATATCCTCCTTAG	construction of C-terminal His12 fused H-NS in K-12(W3110)

2.2.2 ChAP-seq 実験

ChAP(Chromatin Affinity Precipitation) は、好気条件下、37 ℃、50 ml の LB 培地で大腸菌を培養した大腸菌を用いて、オリジナルの手法に則って行った (Ishikawa et al. 2007)。簡潔には以下の通りである。なお、前節で構築した H-NS-12His を発現する K-12 株、SE11 株、SE15 株それぞれの形質転換体について、以下の ChAP-seq 実験をそれぞれ二度ずつ実施した。

$OD_{600} \approx 0.4$ の培養液に、終濃度 1.0%のホルムアルデヒドを加えてタンパク質ー DNA 複合体を固定し、室温で 20 分間静置した。その後、3M Glycine 溶液を加えて (終濃度 90 mM) 5 分間静置した。固定後、遠心分離によって菌体を回収し、TBS バッファで洗浄して-80 ℃で保存した。氷上にて、3 mL UT バッファ(100 mM HEPES, 50 mM imidazole, 8 M urea, 0.5 M NaCl, 1% Triton X-100, 10 mM β-mercaptoethanol, 1 mM PMSF, pH 7.4) 内で超音波処理を施し、その後、細胞片を取り除くために 14,000 rpm で 10 分間遠心分離を行った。上清のうち 100 ー 200 μl を WCE(whole cell extract) 画分として分離し、残った上清を、Dynabeads Talon (Invitrogen, Norway) 50 μl と混合した。混合溶液は、4 ℃で一晩回転させた。Talon は UT バッファで 5 回洗浄し、タンパク質ー DNA 複合体は、溶出バッファ(100 mM Tris-HCl, pH 7.5, 0.5 M imidazole, 1% SDS, 10 mM DTT) 400 μl を用いて溶出させた。 溶出液を Microcon-100 (Millipore, Ireland) に通して、非特異的に結合しており架橋されていない 100 kDa 以下のタンパク質を取り除いた。膜上に保持されたタンパク質ー DNA 複合体を 3 回、洗浄バッファ(100 mM Tris-HCl, pH 7.5, 1% SDS, 10 mM DTT) を用いて洗浄し、50 μl のバッファを加えることで回収した (ChAP 画分)。ChAP 画分、WCE 画分それぞれにおける、DNA とタンパク質間の架橋は、65 ℃ で一晩加熱することによって分離した。DNA は Qiaquick (QIAGEN, Germany) を用いて精製した。H-NS-His12 の回収効率は、抗 His タグ抗体を用いたウェスタンブロッティングによって確認した。

シーケンシングは、Illumina GA (Illumina, USA) を用いて行った。SE11, SE15, K-12 の 3 株それぞれの ChAP 画分、WCE 画分 (Whole Cell Extract : H-NS の結合の有無と無関係に回収したゲノム DNA 断片であり、コントロールとして用いられる)、また、各株の再現性実験を含めた計 6 サンプルについて、36 bp のシングルエンド法でシーケンシングを行った。ライブラリの構築と PCR 増幅は Illumina のプロトコルに従った。1 回目の実験からは、各サンプルそれぞれ約 800 万から 1100

万本のリードを、2 回目の実験からは、各サンプルそれぞれ約 500 万から 1000 万本のリードを得た。

2.2.3 シーケンシングリードのマッピングおよびデータの正規化

Illumina GA から得られた 36 bp のリードを、各株のリファレンスゲノム配列にマッピングした。使用したツールは、アラインメントツールの BLAT (Kent 2002) で、

1. アラインメント中にギャップを含まない
2. ミスマッチは 2 bp まで

の 2 条件のもと、マッピングを行った。複数箇所にヒットしたリードは扱わず、1 箇所にユニークにヒットしたリードのみを以下の解析に用いた。また、大腸菌には rRNA オペロンが 7 コピー存在するため、36 bp という短いリードが本来 7 領域のどこにあったのか、区別してマッピングすることは難しいことを考慮して、rRNA オペロン上の結合強度はゼロと定義し、以後の解析ではこの領域の値を無視した。

Illumina シーケンシングにかける DNA ライブラリは長さ 200 bp の DNA 断片であるため、BLAT で報告されたマッピング位置から、リードは 3 ’ 側へ 200 bp 延長してマッピングした。ゲノム上の平均リードカバレッジは、WCE サンプルで x 96 から x 243、ChAP サンプルで x 52 から x 249 であった。

3 株およびそれぞれの再現性実験で得られたデータを比較するため、各サンプルの結合強度データにグローバルスケーリングによる正規化処理を施した。3 株の ChAP、WCE、およびそれぞれの再現性実験の計 12 サンプルに関して、ゲノム上 1 塩基ごとに、「その塩基位置でマッピングされたリードカバレッジを、そのサンプルにおけるゲノム全体のリードカバレッジの中央値で割り算した値」を各サンプルの正規化結合強度として割り当てた。その後、ゲノム上のある塩基位置の H-NS 結合強度を、「その位置における正規化 ChAP 結合強度を、同じ位置における正規化 WCE 結合強度で割った値」として決定した。もし、ある位置における正規化 ChAP 結合強度あるいは正規化 WCE 結合強度のどちらかがゼロ (たとえばゲノム上で重複した領域など) だったならば、その位置の H-NS 結合強度はゼロとした。結合強度の値は 0 から約 8,000 の広い範囲に分布していたため、以下の解析では 10 を底とした対数をとった値を結合強度として用いた。図 2.2 は、各サンプルにおける処理のフローチャートである。

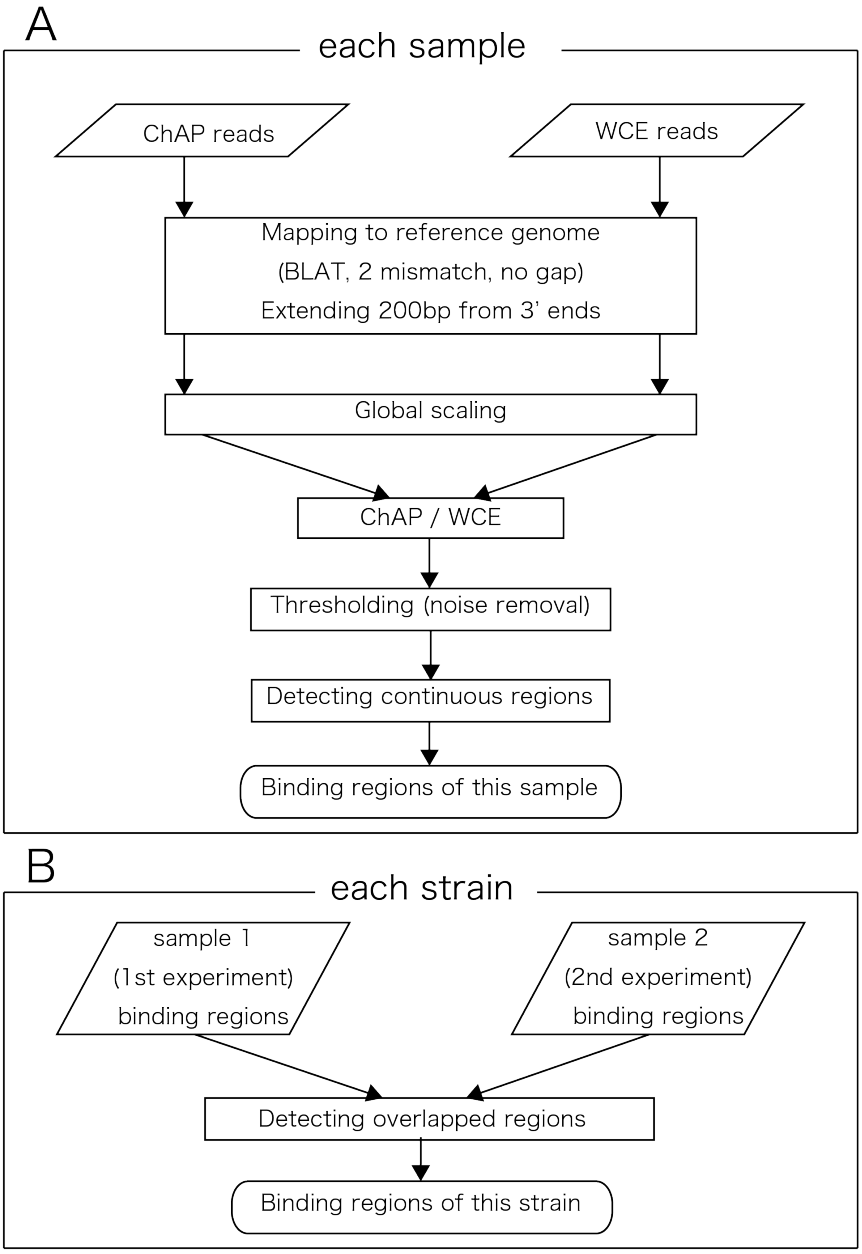

図 2.2: (A) 大腸菌 3 株の ChAP サンプル、WCE サンプル、およびそれぞれの株の再現性実験によって得られた、計 12 サンプルに対する処理内容。これによって、大腸菌 3 株それぞれで 2 種類ずつ（1st experiment, 2nd experiment）の「暫定 H-NS 結合領域」が決定される。(B) 各株について、2 実験の「暫定 H-NS 結合領域」から、「真の H-NS 結合領域」を決定する。

正規化処理を施した結果、株間で、相同な位置がほぼ同じ結合強度の値をとっているかを確認するために、次節で定義する“Common”領域上で 200 bp のウィンドウを 100 bp ステップでスライドさせ、各ウィンドウ内の平均結合強度をプロットした (図 2.3A)。相関係数の値は、いずれの株ペアでも非常に高かった (pearson's r ＞ 0.69)。また、各株のゲノム上の同一の位置で、再現性実験の結果がほぼ同じ結合強度の値をとっているかを確認するため、各株の全ゲノム上で、同一の方法で 2 実験間の結合強度を比較した (図 2.3B)。いずれの株でも、2 実験間で相関係数は 0.8 以上となった。2 実験間で H-NS 結合強度が完全に一致するほどの再現性は得られなかったが、この結果に関しては以下の２点に留意する必要がある。まず、2 回目の実験では 1 回目の実験と比較して得られたリード数が少ないことにより、1 回目の実験で高い結合強度が観測された領域において、2 回目の実験では結合強度が低く観測されている可能性がある。また、図 2.3B はゲノム全域における結合強度の比較であるため、次節で定義する「ノイズ」と見なされ得る領域、すなわち結合強度値の変動が実験的な誤差と見なされ得る領域も含まれた比較となっている。これらの要因があるにも関わらず、2 実験間の正規化結合強度値は非常に高く相関していた。したがって、ChAP-seq 実験の再現性の高さが示されるとともに、上記の処理によってデータの正規化が正しくなされていると判断し、以下の解析にはこの正規化 H-NS 結合強度を使用した。

2.2.4 H-NS 結合領域の決定

H-NS 結合強度の分布は、値 1.0 を中心としてほぼ正規分布する“ノイズ”成分と、10.0 から 8000.0 の広い範囲に分布する“シグナル”成分からなる、二峰性の分布を示していた。ノイズ成分の分布は、平均値 μ を分布の最頻値とし、標準偏差 σ を 0.2 とした正規分布によって近似できる（図 2.4)。H-NS 結合領域を決定するにあたって、ノイズ成分を取り除くために、μ を各サンプルの分布の最頻値、σ を 0.2 として、$Z = \mu + 3\sigma$ を閾値として設定した。まず、「暫定 H-NS 結合領域」を以下の 3 ステップで決定した。(各株における H-NS 結合領域同定のフローチャートは図 2.2B)

1. 各サンプル (SE11, SE15, K-12 の 3 株のそれぞれ 2 実験、計 6 サンプル) において、閾値以上の値を示しているゲノム上の位置を検出する。

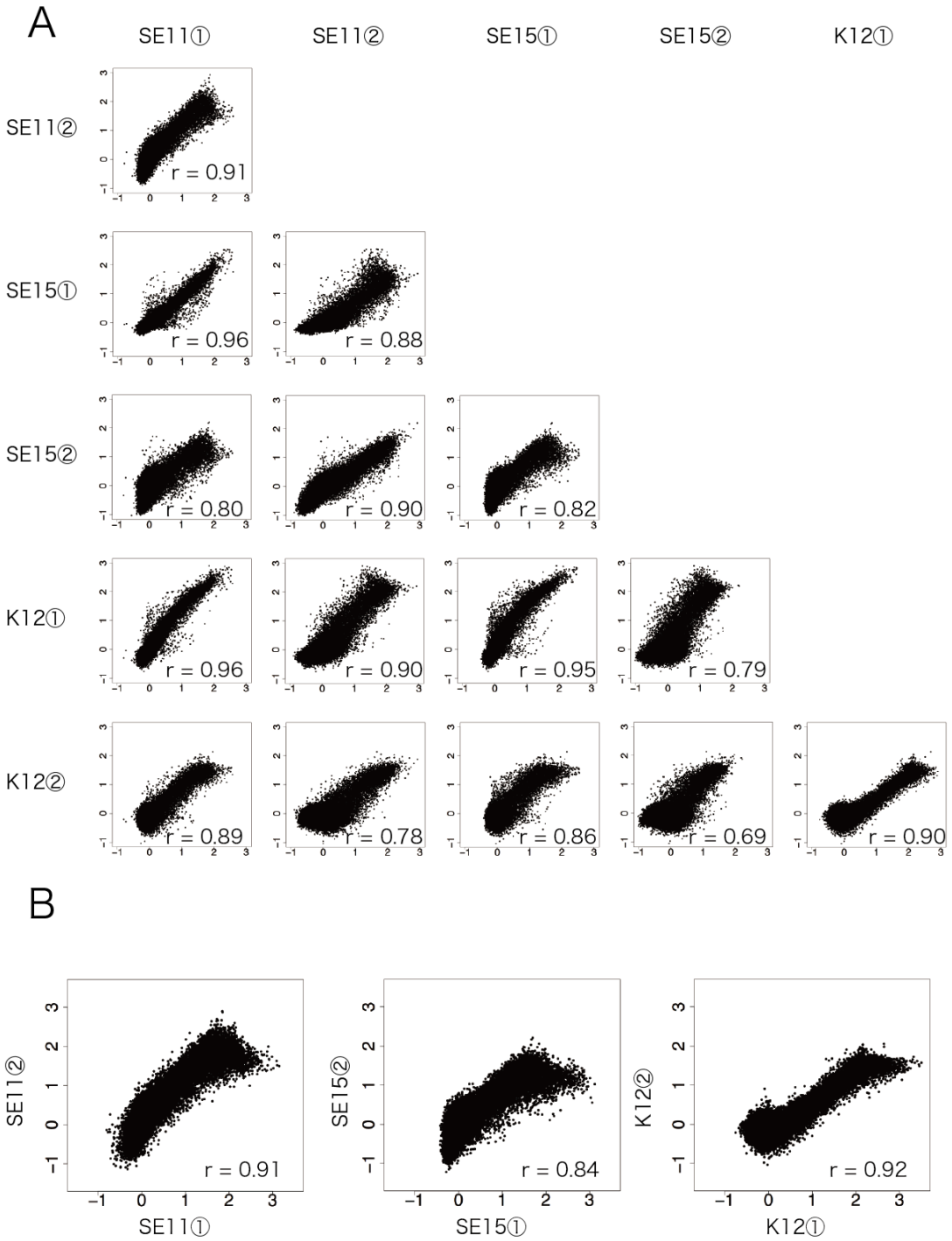

図 2.3: (A) “Common” ゲノム領域における、大腸菌 3 株およびそれぞれの再現性実験（1st experiment, 2nd experiment）の正規化結合強度の比較　(B) 全ゲノム上における、各株の 2 実験の正規化 H-NS 結合強度の比較

2. 閾値以上の値をとっている 2 つの領域の間の、閾値以下の領域の長さが 200 bp 以下の場合、この 2 つの H-NS 結合領域は連結し、その間の閾値以下の領域も H-NS 結合領域とみなす。

3. 上記 2 ステップを終えた段階で、長さが 200 bp 以下の H-NS 結合領域は、偽陽性の H-NS 結合領域として、結合領域のリストから取り除く。

この 3 ステップによって、6 サンプルそれぞれで、「暫定 H-NS 結合領域」が決定される。最後に、同一株の 2 実験を比較し、ゲノム上でオーバーラップしている H-NS 結合領域を「真の H-NS 結合領域」として決定した。以上の手順 によって、3 種の大腸菌のゲノム上 H-NS 結合領域を決定した。以下、「結合領域」という場合、この手順によって得られた「真の H-NS 結合領域」を指すこととする。

本研究の ChAP-seq 実験によって取得したリードは 36bp のシングルエンドリードであり、配列の短さに起因する問題のために、リファレンスゲノムへの BLAT を用いたマッピングが正確に行われなかったリードが多く含まれてしまっている可能性はある。上記の手順によって決定した H-NS 結合領域の正確性を検証するため、異なるショートリードマッピングツールを用いて決定した H-NS 結合領域との比較を行った。検証用のマッピングツールとして Bowtie 2 (Langmead and Salzberg 2012) を使用した。マッピング後、H-NS 結合領域を決定するまでの解析は同様に行った。BLAT を用いたマッピングによって決定した H-NS 結合領域と、Bowtie 2 を用いたマッピングによって決定した H-NS 結合領域を比較した結果、いずれの株においてもそれらの H-NS 結合領域は 97%以上オーバーラップしていた。したがって、上記の手順で決定された H-NS 結合領域は、マッピングツールによらず再現性が高いことを確認した。以後の解析では、BLAT によってマッピングした結果に基づく H-NS 結合領域を用いる。

2.2.5 大腸菌 3 株のゲノム上の “ Common ” 領域の決定

本研究で対象とした大腸菌は、異なった系統群に属する 3 種の大腸菌であるため、それぞれゲノム構造が異なる。本研究では、各株ゲノム上の H-NS 結合領域を比較解析することを目的としているため、それに先立って、全ゲノムアラインメントによるゲノム構造の比較を行った。全ゲノムアラインメントは、アラインメント

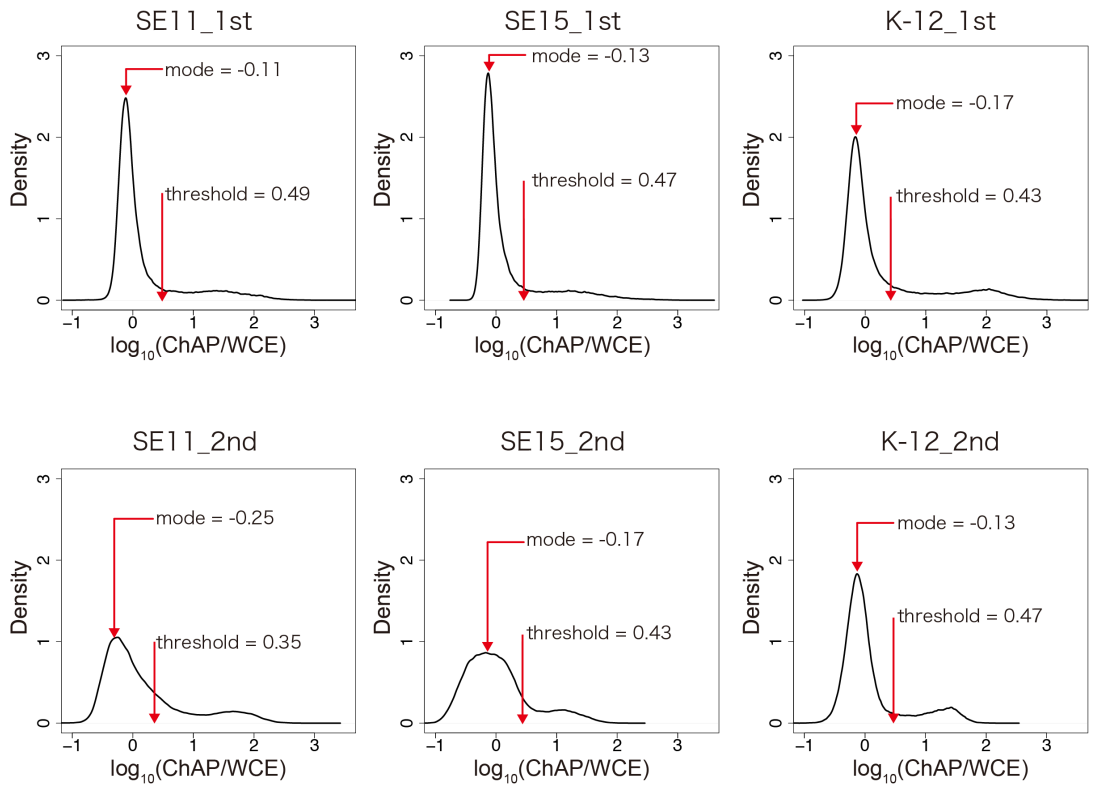

図 2.4: 横軸は H-NS 結合強度（正規化 ChAP/正規化 WCE）について 10 を底とした対数をとった値、縦軸はそれらのゲノム上頻度。大腸菌 3 株それぞれの 2 実験、計 6 サンプルについて、黒色の線は、ゲノム上の密度推定曲線を示している。結合強度は、値ゼロを中心とした“ノイズ”成分と、広い値に分布する“シグナル”成分からなる。密度推定曲線の最頻値を平均値（μ）とし、0.2 を標準偏差（σ）とした正規分布は、各サンプルの“ノイズ”成分とよく一致する。そこで、$\mu + 3\sigma$ を各サンプルの閾値として設定し、ノイズを除去した。

ツールの Mauve 2.3.1(Darling et al. 2004, 2010) を用いて行い、Progressive Mauve algorithm をデフォルトのパラメータ設定で使用して、SE11 株、SE15 株、K-12 株のゲノム配列の全ゲノムアラインメントを得た（図 2.5）。各株のゲノム配列および遺伝子アノテーション情報は、NCBI の GenBank データベースからダウンロードした (SE11：AP009240.1, SE15：AP009378.1, K-12(W3110)：AP009048.1)。

全ゲノムアラインメントの結果から、大腸菌 3 株が共通して保持しているゲノム領域を、“Common”領域として決定した。“Common”領域は、

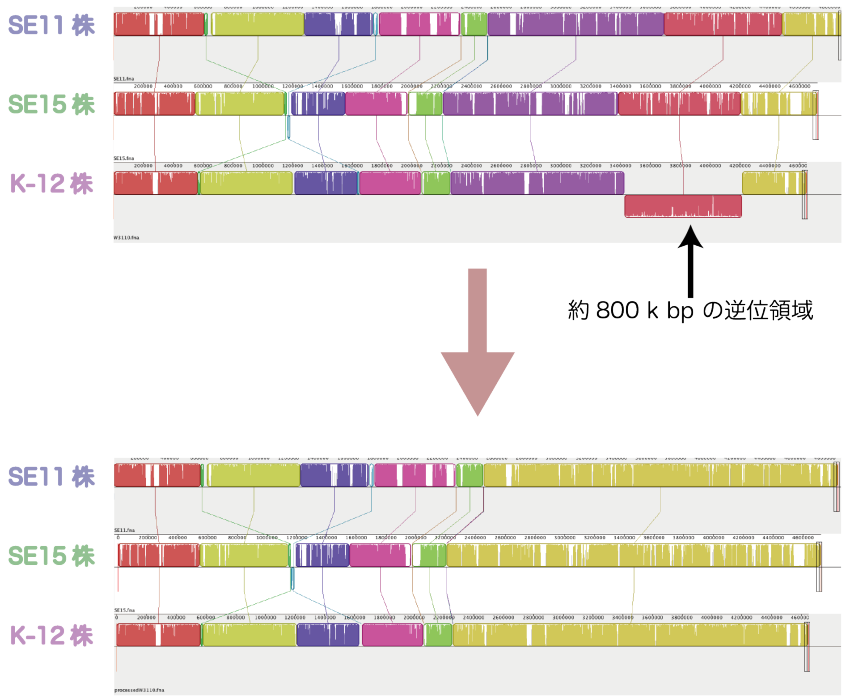

図 2.5: K-12 株のゲノムにおいて約 800 kbp のゲノムリアレンジメントが生じていた。この逆位領域の配列を手動で補正し、再度 Mauve にかけることによって、3 株の全ゲノムアラインメントを得た。

1. 3 株のゲノム配列がすべてアラインされている
2. シンテニーが保存されている

ことを条件とした。シンテニーの保存を条件としているため、この結果得られた“Common”領域は、3 株の大腸菌が分岐する以前の祖先種のときから維持され続けたゲノム領域と考えることができる。

Mauve プログラムは、.backbone ファイルとして、全ゲノムアラインメント結果を、株間の DNA 配列保存度合いによって領域として区分けし出力する。.backbone ファイルの情報に従って、各株のゲノム配列について、

1. 3 株すべてが保持している“Common”領域
2. 3 株中 2 株に存在し、1 株には存在しない“Shared”領域
3. 他の株と相同ではない、1 株が特異的に保持している“Specific”領域

を決定した。最後に、各株のゲノム配列上の“Common”領域を連結して、仮想的な“Common”染色体を構築した。この“Common”染色体は、3,888,365 bp の長さとなった。実際には、全ゲノムアラインメントの結果、それぞれの株のゲノム配列にはわずかなギャップが挿入されるので、各株の“Common”領域の長さは、これよりわずかに短くなる (SE11 : 3,886,369 bp, SE15 : 3,886,157 bp, K-12 : 3,886,242 bp)。全ゲノムアラインメントの際、K-12 株のゲノム配列に、約 800 kbp の長い領域の逆位が見出された (2 つの rRNA オペロンに挟まれた 3,423,096 - 4,216,800 bp の領域に起きたゲノムリアレンジメント)。この領域は他の 2 株と相同な領域を多く含むが、“シンテニー保存”の条件を満たさないため、上記の手順では“Common”領域からは除かれてしまう。それを回避するため、アラインメントに先立って、K-12 株のこの領域は手動で逆転させ、 その上で上記の手順を行った（図 2.5)。この他に報告された逆位や転座の領域は、それが 3 株すべてで相同な領域であっても、“Common”領域に含めなかった。ここで定義した“Common”領域は以上の条件を満たしているため、3 株の種分岐以前からそれぞれのゲノム上に保持され続け、その後に生じた相同組換えや水平伝播の影響からは免れている領域であると考えられる。

2.3 結果

2.3.1 ChAP-seq 解析による大腸菌 3 株の H-NS 結合領域の比較

大腸菌 3 株 (SE11 株、SE15 株、K-12 株) について、それぞれ 2 回ずつ独立に ChAP-seq 解析を行い、各株のゲノム上における H-NS 結合領域を決定した。各株 2 回の実験の再現性は高く、各株ゲノム上の H-NS 結合領域分布は安定していることを確認した（図 2.6)。

3 株の H-NS 結合領域は、これら 2 回の実験でオーバーラップしている領域として決定した。H-NS 結合領域はそれぞれ、SE11 株ゲノム上の 822,561 bp、SE15 株ゲノム上の 642,859 bp、K-12 株ゲノム上の 697,762 bp をカバーしていた（表 2.2)。これらの H-NS 結合領域はそれぞれの株のゲノム全体のうち、13 - 17%に相当する。

先行研究によって、H-NS は水平伝播によって獲得された領域に優先的に結合し、それらの領域上に存在する外来性遺伝子の発現を抑制することが知られている (Lucchini et al. 2006, Navarre et al. 2006, Oshima et al. 2006, Navarre et al. 2007, Doyle et al. 2007, Dorman 2007)。そこで、アラインメントツールの Mauve

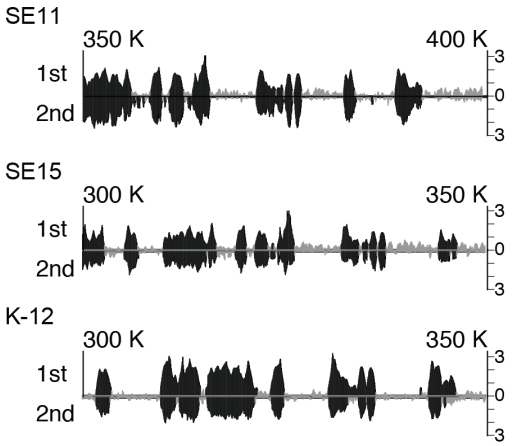

図 2.6: 上から SE11 株 (350Kbp - 400Kbp)、SE15 株 (300Kbp - 350Kbp)、K-12 株 (300Kbp - 350Kbp) のゲノム上 H-NS 結合領域分布の一部を示している。上向きの棒グラフで示しているのが 1 回目の ChAP-seq 実験によって得られた H-NS 結合強度であり、下向きの棒グラフで示しているのが 2 回目の ChAP-seq 実験で得られた H-NS 結合強度である。黒色が“シグナル”として検出された結合強度であり、灰色が“ノイズ”として検出された結合強度を示している。

によって決定された、各株の“Common”領域 (3 株すべてが共通して保持している領域)、“Shared”領域 (3 株中 2 株が保持している領域)、“Specific”領域 (それぞれの株にユニークな領域) 上の H-NS 結合領域の分布を調べた。ゲノム構成は 3 株とも類似しており、ゲノムの 80 - 84%が“Common”領域、それぞれのゲノムの 9%(SE11), 3%(SE15), 10% (K-12) が、“Shared”領域、11%(SE11), 15%(SE15), 6%(K-12) が“Specific”領域であった（表 2.2）。各領域における H-NS 結合領域の割合を調べた結果、いずれの株においても、“Shared”または“Specific”領域は、その領域内の 30 - 38%が H-NS 結合領域に覆われていることがわかった。一方、“Common”領域は 10 - 11%程度が H-NS 結合領域に覆われていた（表 2.2 太字の項目）。3 株における“Shared”領域、“Specific”領域上の H-NS 結合領域割合それぞれについて、“Common”領域上の H-NS 結合領域割合との独立性をカイ二乗検定によって検定した結果、いずれにおいても「割合に差がない」という帰無仮説は棄却された（$p\text{-}value < 0.001$）。

したがって、各株のゲノム上で多様化している“Shared”領域や“Specific”領域は、株間で共有されている“Common”領域よりも、H-NS に覆われやすい傾向を持つことが示された。

表 2.2: 各株の H-NS 結合領域のサイズ

	SE11	SE15	K-12
ゲノムサイズ (bp)	4,887,515	4,717,338	4,646,332
“ Common ” 領域の長さ (bp)	3,886,369	3,886,157	3,886,242
ゲノム中の “ Common ” 領域の割合（%）	79.52	82.38	83.64
“ Shared ” 領域の長さ (bp)	444,371	139,887	479,569
ゲノム中の “ Shared ” 領域の割合（%）	9.09	2.97	10.32
“ Specific ” 領域の長さ (bp)	556,775	691,294	280,521
ゲノム中の “ Specific ” 領域の割合（%）	11.39	14.65	6.04
H-NS 結合領域の長さ (bp)	802,561	642,859	697,762
ゲノム中の H-NS 結合領域の割合（%）	16.4	13.6	15.0
H-NS 結合領域の数	506	510	436
H-NS 結合領域の長さの平均値 (bp)	1611	1514	1789
“ Common ” 領域中の H-NS 結合領域 (bp)	451,643	383,226	427,731
” Common ” 領域中の割合（%）	**11.6**	**9.9**	**11.0**
“ Shared ” 領域中の H-NS 結合領域 (bp)	170,778	50,917	181,704
“ Shared ” 領域中の割合（%）	**38.4**	**36.4**	**37.9**
“ Specific ” 領域中の H-NS 結合領域 (bp)	180,140	208,716	88,327
“ Specific ” 領域中の割合（%）	**32.4**	**30.2**	**31.5**

2.3.2 “ Common ” 領域上の H-NS 結合は株間できわめて精確に保存されている

“ Shared ” や “ Specific ” 領域は、“ Common ” 領域と比較して、より H-NS に覆われやすいという傾向が見られたものの、H-NS 結合領域の 50%以上は、 3 株が共通して保持する “ Common ” 領域上に存在していた（表 2.2）。“ Common ” 領域上の H-NS 結合領域の長さは、3 株で類似している (SE11: 451,643 bp, SE15: 383,226 bp, K-12: 427,731 bp)。これは、“ Common ” 領域内における H-NS の結合が高度に保存されていることを示唆している。図 2.7 に、“ Common ” 領域上の

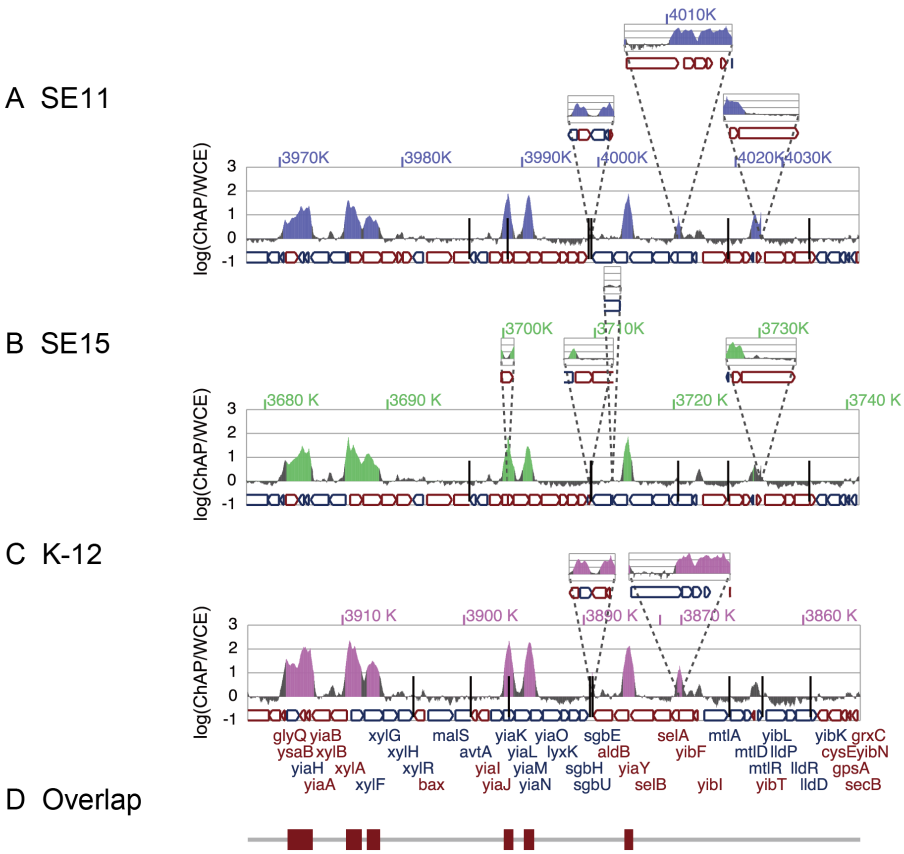

図 2.7: 上から SE11 株 (A)、SE15 株 (B)、K-12 株 (C) の “ Common ” 領域の一部を示している。各株で挿入・欠失している “ Shared ” 領域や “ Specific ” 領域は、“ Common ” 領域である 3 株の水平ラインのそれぞれから、三角形に伸びた領域として図示されている。それぞれの株の H-NS 結合強度波形について、色のついた部分が H-NS 結合領域、灰色の部分が閾値以下のノイズを示している。“ Overlap ” のライン (D) は、“ Common ” 領域において、3 株すべての H-NS 結合領域がオーバーラップしている部分を、赤いボックスで示している。

3 株の H-NS 結合分布の一部を示す。図 2.7 から、“ Shared ” や “ Specific ” 領域 (図 2.7 で、“ Common ” ゲノムとしてアラインされている水平のライン上の所々に、各株で挿入されている三角形の領域) で H-NS の結合分布が多様化している一方、“ Common ” 領域上では H-NS の結合が非常に高度 に保存されていることがわかる。

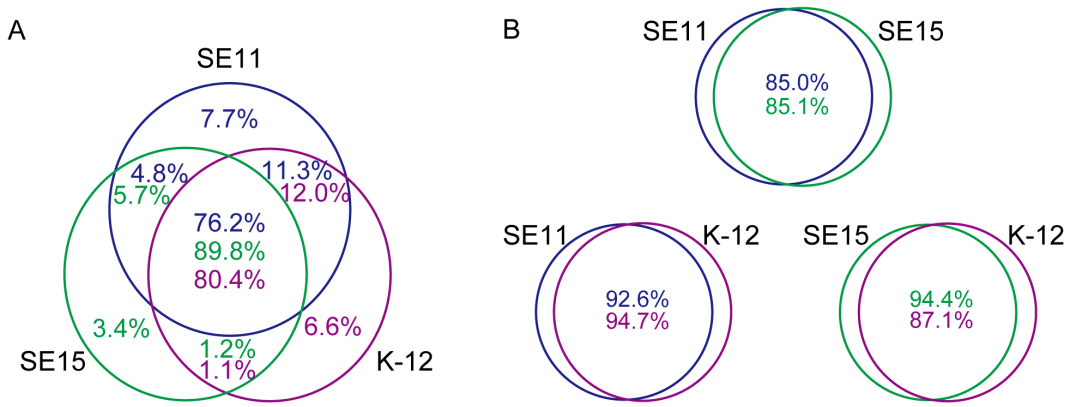

図 2.8: (A) “ Common ” 領域における H-NS 結合領域のオーバーラップ
(B) “ Shared ” 領域における H-NS 結合領域のオーバーラップ

そこで、“Common”領域上 H-NS 結合領域が株間でどの程度の長さオーバーラップしているかを調べた。その結果、各株の “ Common ” 領域上 H-NS 結合領域のうち、76.2%(SE11), 89.8%(SE15), 80.4%(K-12) もの割合が、3 株でオーバーラップしており、各株がユニークに持つ “ Common ” 領域上 H-NS 結合領域はわずか 3.4 - 7.7%であった（図 2.8)。したがって、“ Common ” 領域上の H-NS の結合は、3 株で高度に保存されていることが明らかとなった。これは、H-NS が近年の水平伝播に関わるだけではなく、3 株の種分岐以前から存在するゲノム領域上においても、長期的に H-NS の結合が維持される場合があるということを意味している。また、“Shared”領域に関して、株間の H-NS 結合領域オーバーラップを調べた結果、“ Shared ” 領域に関しても H-NS の結合が非常に高度に保存されていることが明らかとなった (85.0 - 94.7%, 図 2.8)。

2.3.3 H-NS 結合領域の局所的保存性

ゲノム構造の変化と、H-NS 結合領域の変化との関係性をより詳細に検討するために、ドットマトリクス法を利用した配列解析を行った。その結果、各株の H-NS 結合領域は、局所的に見ても一致しているだけでなく、それぞれの領域の長さも、株間でよく保存されていることがわかった（図 2.9A)。さらに、H-NS 結合領域は、その周辺のゲノム配列に逆位が生じたとしても、大きくは変化しないことが明らかとなった（図 2.9B、赤い矢印で示した領域)。

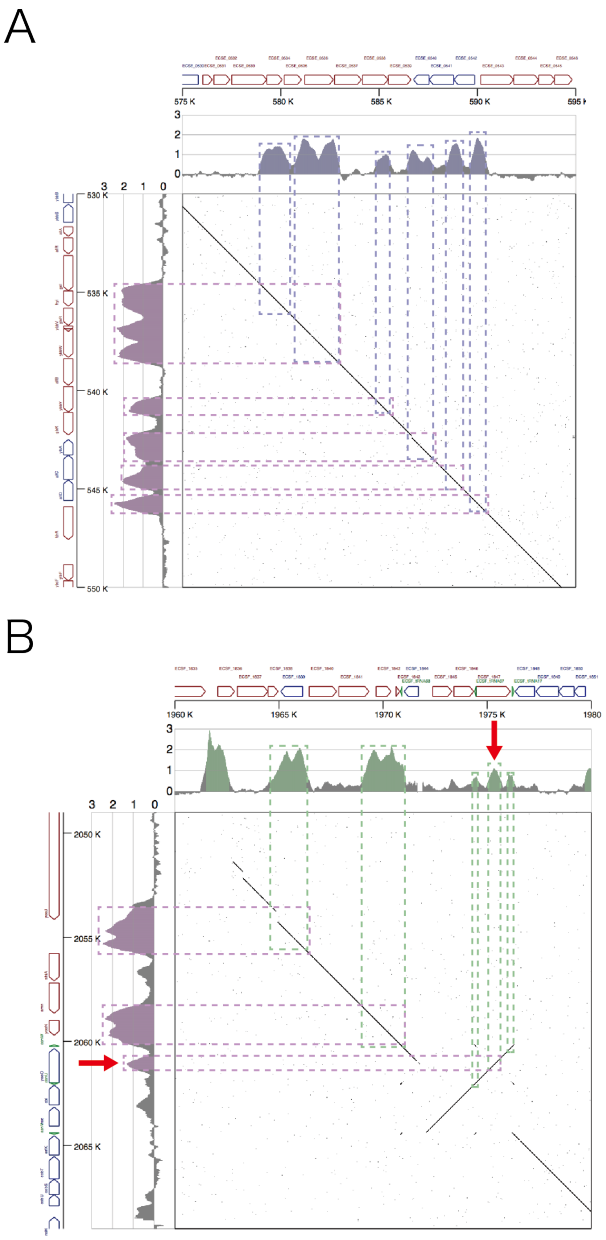

図 2.9: (A) SE11 株と K-12 株の比較 (B) SE15 株と K-12 株の比較。一部に配列逆位領域を含んでいるが、その領域上の H-NS の結合位置（赤矢印）が一致している。

2.4 考察

大腸菌 3 株のゲノムの“Common”領域、“Shared”領域、“Specific”領域それぞれにおける H-NS 結合領域の分布を調べた結果、H-NS に覆われる割合は、“Shared”領域および“Specific”領域において高いことがわかった。“Shared” や“Specific”領域は、3 株の種分岐後、比較的最近の組み換えや水平伝播によって、獲得・欠失された領域であると考えられるため、この結果は、H-NS が外来性遺伝子の発現を抑制するために、外来性の配列に結合しやすい傾向があるという仮説を支持している。

一方、3 株が長期的に保持していると考えられる“Common”領域においても、H-NS の結合領域は多数存在し、それらの H-NS 結合領域はきわめて精確に保存されていることがわかった。大腸菌 3 株の“Common”領域における H-NS 結合領域の 80%以上はオーバーラップしていた。各株の結合強度閾値の設定によって、いくつか弱い H-NS 結合領域が解析対象から除外されているので、実際にはこの値はさらに高い可能性がある。また、“Common”領域における H-NS 結合領域の長さが株間でわずかに異なる点も、多くの場合、領域周辺をどこまで H-NS 結合領域として含めるか、各株による結合強度閾値の設定による影響と考えられる。したがって、“Common”領域上のほとんどすべての H-NS 結合領域で、H-NS の結合がきわめて高度に保存されていると結論できる。

また、3 株中 2 株で共通して保持する“Shared”領域においても、その領域内の H-NS の結合が高度に保存されていることが見出された。大腸菌 3 株の系統関係 (図 2.1) から、“Shared”領域は、

1. 3 株の共通祖先種の時代から存在するが、いずれか 1 株で欠失した領域
2. 3 株の共通祖先種から SE15 株が分岐後、SE11 株と K-12 株の祖先種で水平伝播によって獲得された領域
3. 3 株中 2 株で独立に、相同な DNA 配列の水平伝播が生じて獲得された領域

のいずれかである。1 であれば、“Common”領域の場合と同様に、H-NS 結合の長期的な維持を意味し、2 または 3 であれば、水平伝播によって獲得されたゲノム領域における H-NS の結合分布が、大腸菌株間で高度に一致していることを示している。

特筆すべきは、これらの H-NS の結合維持において、それぞれの結合領域の長さも維持されているという点にある。H-NS 結合領域は平均して 1,500 - 1,800 bp の長さであるが、領域ごとに数百塩基から数万塩基にいたるまで多様である。様々な長さの H-NS 結合領域が、“Common”領域上では、 その長さを保ったまま、維持されている (図 2.9A)。さらに、H-NS 結合領域周辺で、配列の挿入・欠失や逆位が起きたとしても、局所的な H-NS 結合領域はその周辺の変化の影響を受けず、維持されている (図 2.7, 図 2.9B)。

以上の結果から、H-NS は決してゲノム上にランダムに結合するのではなく、また、周辺の H-NS 結合領域との距離や分布の間隔など、ゲノム上のマクロな関係性で維持されるわけではないことがわかった。したがって、本章で観察された“Common”領域および“Shared”領域における H-NS 結合の維持は明らかに、局所的な DNA の一次配列が保存されることによって達成されている。そしてこのような DNA 配列は、H-NS の結合領域に特異的であると共に、それぞれの H-NS 結合領域の長さをも特徴づける何らかのパターンを有しており、さらに、それらの配列が大腸菌 3 株の“Common”領域上において進化的なバイアスを受け、長期的に維持されていると考えられる。

第3章　H-NS結合配列の多様化

本章では、H-NS結合領域下のDNA配列について大腸菌系統群における分子進化解析を行うことにより、Dormanが提唱した仮説の実証を試みる。比較する大腸菌株を44株に広げ、オーソログ遺伝子、および本研究で新たに定義する相同遺伝子間領域の進化距離の比較から、H-NS結合領域では進化が加速していることをみる。さらに、相同遺伝子間領域について遺伝子の転写方向を考慮した分類をすることによって、そのような進化の加速が、H-NSによる転写抑制を介した間接的な影響、すなわちDNA配列にかかる選択圧の緩和によって生じていることを示す。

3.1　緒言

前章での解析の結果、大腸菌3株の“Common”ゲノム領域上で、H-NSの結合位置・結合強度・結合の広がりが高度に維持され、それは基本的に、DNA配列の保存によって達成されていることが示唆された。なんらかの特徴が生物間で共有され、進化の乱雑さに抗って維持されているとき、特にそれが大腸菌など世代時間の短い生物種で生じているときには、そこに生物学的な必然性を仮定するのは自然である。H-NSの結合はなぜ長期的に維持されているのか。次の2つの可能性が考えられる。

1. H-NSがある位置で結合しているという情報が、ゲノムの高次構造形成にとって重要である
2. H-NSがある位置で結合しているという情報が、その位置にコードされている情報の管理にとって重要である

前者は構造的条件であり、後者は指示的条件である。

構造的条件については、細菌ゲノムの高次構造が従来考えられていたよりも精緻に組織化されていること、また、H-NSをはじめとした核様体タンパク質群の結

合が、DNA にそれぞれ特有の屈曲構造をもたらすことによってそのような組織化に寄与している可能性があることなどにより、近年盛んに議論されている (Dame 2005, Maurer et al. 2009, Dame et al. 2011, Wang et al. 2011, Dorman 2013)。だが、本研究では H-NS の結合がもたらすゲノムの高次構造については深く議論しない。前章でみたように、H-NS の結合は数十 bp の解像度で高度に維持されていた。そのような結合領域の維持が、染色体のマクロな立体構造の維持のためになされているとは思えない。たしかに細菌のゲノムには、Ori-Ter 対称性や gene strand bias, GC skew をはじめとした、大局的な構造ルールが存在する (Rocha 2004, Dorman 2013)。しかし、図 2.7 を見てもわかるように、大腸菌のゲノムには数千 bp サイズの DNA 配列の挿入・欠失が頻繁に生じている。挿入・欠失によって、H-NS 結合領域間の相対的な位置関係は大きく変化する。このように（相対的位置関係としては）変動し得る領域がゲノムのマクロな高次構造に直接的かつ決定的な影響を与えているとは考えにくい。一方で H-NS が、ゲノムに屈曲を導入するための、だいたいの位置の目安としてのみを目的として結合しているのであれば、これほど精確に結合位置が保存される必要はない。したがって、構造的条件というよりも、H-NS 結合領域下に存在する情報（遺伝子配列、遺伝子制御配列）に影響を与えることにこそ、H-NS 結合の維持の意味があるはずである。

前章で、これまでの報告通り (Lucchini et al. 2006, Navarre et al. 2006, Oshima et al. 2006, Dorman 2007, Navarre et al. 2007)、H-NS はやはり、外来性配列とみなしうる領域に結合しやすい傾向があることが示された。3 株の“Common”ゲノム領域でも多くの H-NS 結合が観察されたが、それらもやはり、3 株の分岐以前の段階ですでに獲得されていた外来性配列に由来する領域である可能性が高い。実際、本章で決定する大腸菌オーソログのうち、H-NS が結合しているものを、複数の論文で提案されている「外来性遺伝子」(Lawrence and Ochman 1998, Garcia-Vallve et al. 2003, Nakamura et al. 2004) と比較した結果、H-NS 結合オーソログは統計的に有意に外来性遺伝子とオーバーラップしていた（$p\text{-}value < 0.001$; フィッシャー正確確率検定）。さらに、公表されている大腸菌の RNA-seq のデータ (Taniguchi et al. 2010) を用いて、本研究で定義した H-NS 結合遺伝子と、H-NS 非結合遺伝子の発現量の分布を比較した結果、H-NS 結合遺伝子は有意に発現量が低いことがわかった（図 3.1）。以上から、先行研究の報告通り、

1. 外来性遺伝子配列に結合する傾向がある

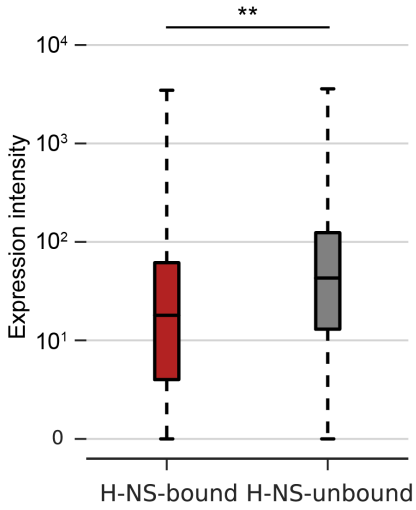

図 3.1: 先行研究 (Taniguchi et al. 2010) の RNA-seq に基づいた大腸菌 K-12 株遺伝子発現量データに関する、H-NS 結合遺伝子（赤）、H-NS 非結合遺伝子（灰）の分布の比較　$^{**}p\text{-}value < 0.001$, wilcoxon rank sum test

2. 結合遺伝子の転写抑制を引き起こす

の二点が、H-NS の大きな特徴であることを確認した。H-NS 結合領域が維持されていることの意味は、これらの状況を踏まえたうえで考える必要がある。

Dorman が提唱した仮説 (Dorman 2007) によれば、H-NS 結合の意味は、悪影響を及ぼし得る外来性遺伝子の転写を抑え、配列の多様化と、その結果として既存の転写調節ネットワークへの統合を促すためであると言われている。この仮説が正しいならば、以下の 2 点が、実際のデータから観察されるはずである。

1. H-NS 結合領域の配列多様性は、H-NS 非結合領域の配列多様性と比較して大きい

2. 配列多様性が増大した結果、H-NS 結合遺伝子をより“巧く”使いこなしている

そこで本章では、上記 2 点を実データから観察することによって、仮説の実証を目指す。大腸菌系統群の配列多様性は、分子進化解析によって評価する。分子進化解析にあたっては、前章で扱った 3 株のみに限定した解析では、解像度として非常に低い。大腸菌系統群は多様なゲノム構造を持っているとはいえ、進化的

には互いに非常に近縁であり、その中でのわずかな配列多様性の差異を検出しなければならない。そのため本章では解析対象を、公共データベースに完全ゲノム情報および遺伝子アノテーション情報が登録されている大腸菌 44 株に広げ、それらについて、オーソログ遺伝子（44 株の分岐以前に共通祖先が有していたと考えられる遺伝子）、および、相同遺伝子間領域（共通祖先が有していた遺伝子“間”領域）を定義する。それぞれについて分子進化解析を実行することで、前者については遺伝子配列そのものの多様性（機能多様性）を、後者については遺伝子制御領域の配列多様性（制御多様性）を評価することができる。

本章の構成は次の通りである。3.2 節では、本研究で対象とした大腸菌株の情報、オーソログの検出と進化距離計算、相同遺伝子間領域の検出と進化距離計算、それぞれの手法について述べる。3.3 節では、これらの結果をまとめる。3.4 節では分子進化解析の結果が、仮説から予想される結果とよく整合していることを示し、とくに H-NS による転写抑制の効果がこれらの結果に決定的な影響を及ぼしていることを議論する。

3.2 材料と手法

3.2.1 大腸菌株

解析に使用する大腸菌は表 3.1 の 44 株である。すべての株のゲノム情報および遺伝子アノテーション情報は RefSeq データベース（NCBI Refrence sequence database）より取得した。

表 3.1: 分子進化解析に使用する大腸菌 44 株

Chromosome ID	#ORFs	Strain name
NC_017626	4800	Escherichia_coli_042_uid161985
NC_008253	4619	Escherichia_coli_536_uid58531
NC_011748	4759	Escherichia_coli_55989_uid59383
NC_017631	4793	Escherichia_coli_ABU_83972_uid161975
NC_008563	4430	Escherichia_coli_APEC_O1_uid58623
NC_010468	4199	Escherichia_coli_ATCC_8739_uid58783
NC_012892	4204	Escherichia_coli_BL21_DE3__uid161949
NC_012967	4204	Escherichia_coli_B_REL606_uid58803
NC_012759	4084	Escherichia_coli_BW2952_uid59391
NC_004431	5369	Escherichia_coli_CFT073_uid57915
NC_017638	4259	Escherichia_coli_DH1_uid162051
NC_009801	4749	Escherichia_coli_E24377A_uid58395
NC_011745	4915	Escherichia_coli_ED1a_uid59379
NC_017633	4697	Escherichia_coli_ETEC_H10407_uid161993
NC_009800	4377	Escherichia_coli_HS_uid58393
NC_011741	4349	Escherichia_coli_IAI1_uid59377
NC_011750	4730	Escherichia_coli_IAI39_uid59381
NC_017628	4757	Escherichia_coli_IHE3034_uid162007
NC_007779	4227	Escherichia_coli_K_12_substr__W3110_uid16351
NC_017660	4697	Escherichia_coli_KO11FL_uid162099
NC_011993	4376	Escherichia_coli_LF82_uid161965
NC_017644	4873	Escherichia_coli_NA114_uid162139
NC_013353	5054	Escherichia_coli_O103_H2_12009_uid41013
NC_013364	4972	Escherichia_coli_O111_H__11128_uid41023
NC_011601	4552	Escherichia_coli_O127_H6_E2348_69_uid59343
NC_011353	5315	Escherichia_coli_O157_H7_EC4115_uid59091
NC_002655	5298	Escherichia_coli_O157_H7_EDL933_uid57831
NC_002695	5230	Escherichia_coli_O157_H7_Sakai_uid57781
NC_013008	5255	Escherichia_coli_O157_H7_TW14359_uid59235
NC_013361	5364	Escherichia_coli_O26_H11_11368_uid41021
NC_017656	4912	Escherichia_coli_O55_H7_RM12579_uid162153
NC_017646	5009	Escherichia_coli_O7_K1_CE10_uid162115
NC_017634	4429	Escherichia_coli_O83_H1_NRG_857C_uid161987
NC_017663	4393	Escherichia_coli_P12b_uid162061

table continued on next page

continued from previous page

Chromosome ID	#ORFs	Strain name
NC_011742	4692	Escherichia_coli_S88_uid62979
NC_011415	4675	Escherichia_coli_SE11_uid59425
NC_013654	4338	Escherichia_coli_SE15_uid161939
NC_010498	4742	Escherichia_coli_SMS_3_5_uid58919
NC_017632	4650	Escherichia_coli_UM146_uid162043
NC_011751	4823	Escherichia_coli_UMN026_uid62981
NC_017641	5117	Escherichia_coli_UMNK88_uid161991
NC_007946	5017	Escherichia_coli_UTI89_uid58541
NC_017664	4608	Escherichia_coli_W_uid162101
NC_017906	5039	Escherichia_coli_Xuzhou21_uid163995

3.2.2 大腸菌オーソログ遺伝子の決定

オーソログ遺伝子を特定するために、まず 44 株すべての遺伝子のアミノ酸配列について、All to All の BLASTP 相同性検索を行った。その後、BLAST 結果を入力として、似た遺伝子のグループを抽出するツールである、OrthoMCL (Li et al. 2003) を用いて、オーソログ候補のクラスタリングを行った。OrthoMCL は、ノードをそれぞれのアミノ酸配列、エッジを配列間の BLAST スコアとしたグラフ構造を考え、そのグラフ上でのランダムウォークを実行して、サブグラフでの滞留時間をもとにしてグラフをクラスタへと分割する、マルコフクラスタリングというアルゴリズムに基づくプログラムである。

大腸菌系統群のオーソログ遺伝子、すなわち大腸菌が様々な株に分岐する以前の共通祖先が持っていた遺伝子を推定するにあたっては、いくつか気をつけなければならない点がある。まず第一に、水平伝播で獲得された遺伝子が含まれてはならない。ある株で祖先系統から保持し続け変異を蓄積している遺伝子が、別の株では水平伝播で獲得されていると、その間の進化距離を過剰に、あるいは過小に見積もってしまう。そのため本研究では、すべてのアミノ酸配列を、ファージ関連遺伝子のデータベースである ACLAME (Leplae et al. 2004)、および IS(insertion sequence) のデータベースである ISFinder (Siguier et al. 2006) のデータセットに対して BLASTP 検索し、ヒットした配列が含まれているオーソログ候補クラスタ

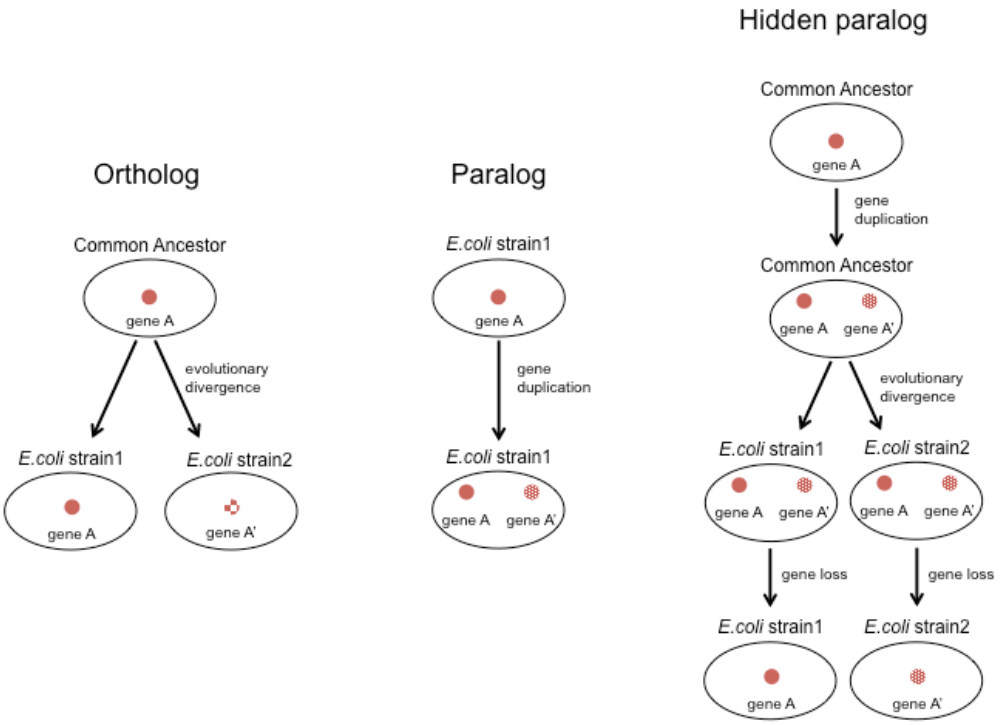

図 3.2: Ortholog, Paralog, Hidden paralog の概念図。Paralog および Hidden paralog は相同な遺伝子クラスタに含まれるが、それらの間の進化距離は、系統間の真の進化距離を反映しない。

をオーソログから除外した。また、進化距離推定の際のバイアスは、Paralog、さらに Hidden Paralog と呼ばれる関係性の遺伝子間で進化距離計算を行ってしまうことによっても生じる（図 3.2）。Hidden paralog は、共通祖先で遺伝子重複が生じていた遺伝子セットがあった場合、分岐後にそれぞれの系統で独立に、どちらかの遺伝子が喪失した結果として現れる (Kuraku 2013)。これらをオーソログであるとみなして進化距離計算を行ってしまうと、共通祖先の時点ですでにで蓄積していた変異のぶん、系統間の実際の距離よりも過大に進化距離を見積もってしまう。このようなバイアスを避けるために、いずれかの株で 2 コピー以上保持されていた遺伝子（遺伝子重複が生じていた可能性のある遺伝子）を含むオーソログ候補クラスタはオーソログから除外した。

44 株すべてに保存されている遺伝子は計 2,135 個であったが、オーソログをこ

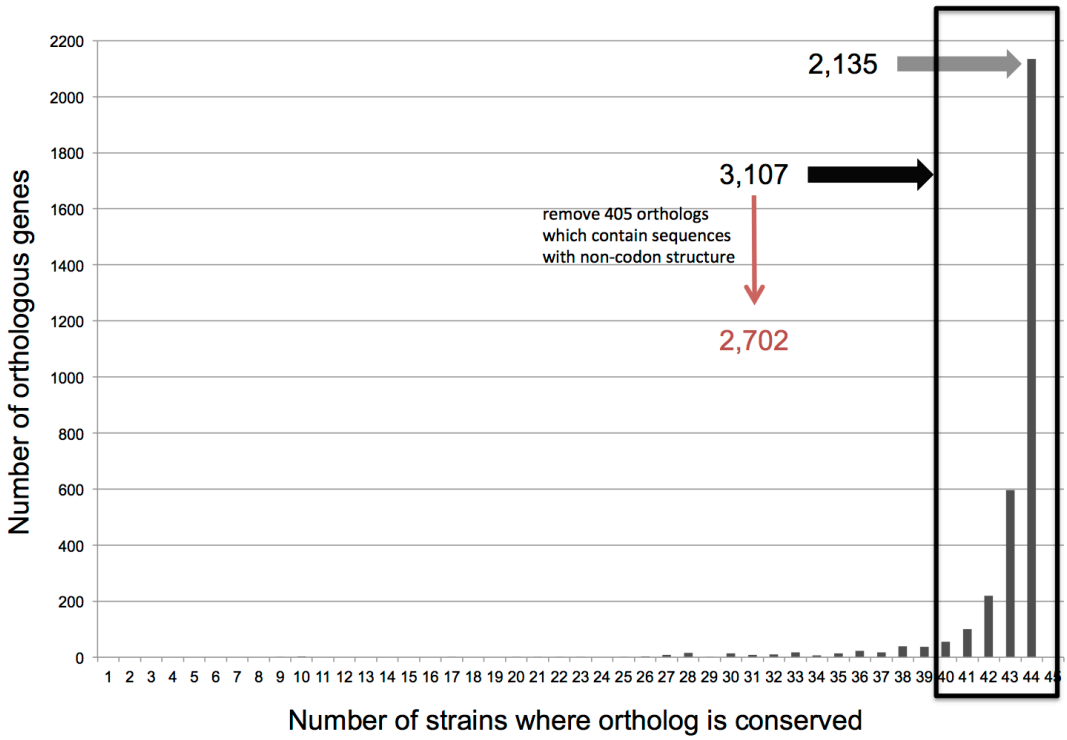

図 3.3: オーソログ候補遺伝子の大腸菌 44 株保存性

れらの遺伝子に限定してしまうと、典型的には 4,000 - 5,000 個の遺伝子を持つ大腸菌という種の、半分以下の遺伝子の進化を扱うことになってしまう。そこで本研究では、44 株のうちの 90%以上（40 株以上）、かつ、K-12 株、SE11 株、SE15 株の 3 株すべてに存在する遺伝子のみを抽出し、3,107 個のオーソログ遺伝子セットを構築した（図 3.3）。

また、そのうちの 405 個では、それぞれに含まれる配列のうち少なくとも 1 配列で、正常ではないコドン構造としてアノテーションされている遺伝子が含まれていた（DNA 配列長が 3 の倍数になっていない、など）。これらは偽遺伝子である可能性があるが、このような配列が含まれていると進化距離が精確に推定できなくなるため、この 405 個はオーソログからは除外した。

以上の操作によって、最終的に 2,702 個のオーソログ遺伝子セットを構築した。

3.2.3 オーソログ遺伝子の分子進化解析

それぞれのオーソログ遺伝子のアミノ酸配列について MAFFT (Katoh and Toh 2008) の G-INS-i アルゴリズムを用いてマルチプルアラインメントを計算し、そのアラインメント構造を維持したまま、それぞれの遺伝子の DNA 配列に変換した。これは、近縁な配列間ではアミノ酸配列レベルでほとんど差異が見いだされないため、DNA 配列レベルでのわずかな変異を検出するために必要な操作である。その後、アラインメント中のギャップ位置や、進化距離推定に不適切なほど多様化してしまっている位置を除去する目的で、それぞれのオーソログ遺伝子のアラインメントに対して、GBLOCKs (Castresana 2000) の codon model を適用した。

オーソログ遺伝子セットの精度評価のため、これらの中からランダムに 100 個の遺伝子を選び、それぞれのアラインメントを結合してひとつの巨大なアラインメントを構築し、その結合アラインメントを用いて、最尤推定法によって 44 株の大腸菌の系統樹トポロジーと枝長の同時推定を行った。使用したツールおよび進化モデルは、PhyML (Guindon and Gascuel 2003) の HKY85 モデル (Hasegawa et al. 1985) である。得られた系統樹（図 2.1）は、先行研究 (Ogura et al. 2009) で報告されている大腸菌株間の系統関係をよく再現しており、これらのオーソログ遺伝子セットの信頼性は高いと考えられる。

すべてのオーソログ遺伝子セットについて、非同義座位の進化距離（Non-synonymous substitution rate; dN）と、同義座位の進化距離（Synonymous substitution rate; dS）を、PAML パッケージ (Yang 1997) の Codeml を用いて推定した（Codeml のパラメータ設定は以下：tree=ML gene tree from PhyML, CodonFreq=F3X4, clock=0, kappa=estimated by ML, omega=estimated by ML, alpha=0, rho=0）。ひとつのオーソログ遺伝子についての大腸菌系統群内非同義座位多様性および同義座位多様性は、各株ペアで計算された dN と dS をそれぞれ、すべての株ペアで平均することによって算出した。

3.2.4 H-NS 結合オーソログ遺伝子の定義の妥当性

H-NS 結合によって遺伝子の進化になんらかの影響があるかを調べるためには、“ H-NS が結合しているオーソログ遺伝子 ” を定義しなければならない。本研究では、前章で見た H-NS 結合位置の株間での高度の保存性を考慮し、そのオー

ソログ遺伝子に対応する K-12 株、SE11 株、SE15 株の遺伝子配列上、少なくとも 1 株で H-NS の結合が見られた場合、そのオーソログ遺伝子は H-NS 結合遺伝子であるとみなす。全株で H-NS 結合が見られない状況は、いずれかの株で結合強度の上昇が見られるものの閾値に達していない場合がほとんどであるので、この定義は問題にならないと考えている。“ H-NS 結合遺伝子 ” の定義をどのように扱うかについては、結果の節でさらに詳しく検討する。

一方で、いくつかのオーソログ遺伝子においては、K-12 株、SE11 株、SE15 株のいずれかに特異的な配列の挿入・欠失があったために、特定の株にのみ H-NS の結合が観察された。その典型的な例は、図 3.4 である。この領域では、*ytfI* という遺伝子が、K-12 株に特異的に挿入されており、それによってその周辺に H-NS の結合が広がっている。K-12 株と近縁な SE11 株では、*ytfI* が存在せず、H-NS 結合領域も存在していない。したがってこの H-NS 結合領域は、K-12 株あるいは K-12 株にきわめて近縁な大腸菌系統群の祖先株における挿入・欠失イベントに付随して生じたものと推測される。すべての H-NS 結合オーソログ遺伝子について、このような「株特異的な挿入・欠失イベントに付随した H-NS 結合領域」が見られるかどうかを精査した結果、全体で 42 個のオーソログ遺伝子が該当した（表 3.2）。これらのオーソログ遺伝子周辺に挿入された配列はごく最近の HGT によって獲得された遺伝子であると考えられ、外来性遺伝子取り込みの初期段階における H-NS の役割を知る上で興味深い事例と考えられる。しかし本研究の目的は、H-NS の長期的な結合によるオーソログ遺伝子の進化への影響であるため、これらを「大腸菌 44 株で H-NS の結合が保存されている H-NS 結合オーソログ遺伝子」として扱うことは適切ではない。したがって以下の解析では、これらの 42 個のオーソログ遺伝子を H-NS 非結合遺伝子として扱った。

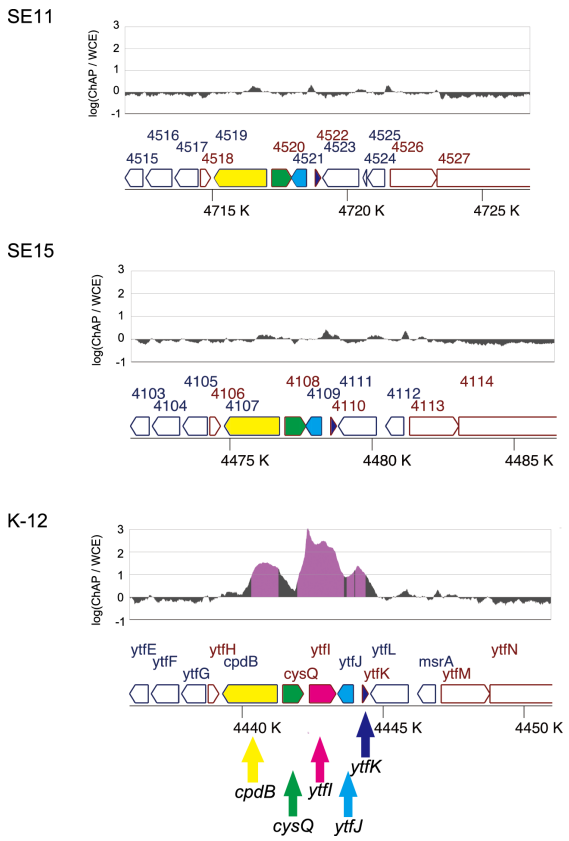

図 3.4: 株特異的な配列の挿入による H-NS の結合

表 3.2: 株特異的な挿入・欠失に付随する H-NS 結合遺伝子

Ortholog name	K-12 insertion	SE11 insertion	SE15 insertion
nhaR	*insA*, *insB*	ECSE_0018 - ECSE_0020	ECSF_0022 - ECSF_0023
yafK	*yafQ*, *dinJ*	ECSE_0246 - ECSE_0247	None
yafL	*yafQ*, *dinJ*	ECSE_0246 - ECSE_0248	None
yajR	None	ECSE_0450 - ECSE_0452	None
cusS	None	ECSE_0626 - ECSE_0633	None
uspG	*ybdR*	ECSE_0676	None
ybgA	*rhsC* - *ybfD*	ECSE_0761 - ECSE_0765	None
phr	*rhsC* - *ybfD*	ECSE_0761 - ECSE_0765	None
trpA	*yciE* - *yciG*	ECSE_1306 - ECSE_1308	None
rluB	None	None	ECSF_1248 - ECSF_1249
btuR	None	None	ECSF_1248 - ECSF_1249
cedA	*ydjO*	ECSE_1899 - ECSE_1901	None

table continued on next page

continued from previous page

Ortholog name	K-12 insertion	SE11 insertion	SE15 insertion
yegK	*yegJ*	None	None
vacJ	*ytdF*	ECSE_2654	None
yfeD	None	None	None
ygaT	*intA* - *yqaD*	ECSE_2904 - ECSE_2911	None
ygiC	None	None	ECSF_2880 - ECSF_2882
exuR	None	ECSE_3375 - ECSE_3378	None
kbl	*htrL*	ECSE_3901	None
pstS	None	ECSE_4015 - ECSE_4018	None
aceK	*arpA*	ECSE_4302	None
iclR	*arpA*	ECSE_4302	None
yjbB	None	None	ECSF_3869
malM	*yjbI*	ECSE_4329 - ECSE_4330	None
yjbN	*yjbL*, *yjbM*	ECSE_4339 - ECSE_4341	None
cysQ	*ytfI*	None	None
rpsT	*insA*, *insB*	ECSE_0018 - ECSE_0020	ECSF_0022 - ECSF_0023
tauA	*yaiO* - *yaiS*	ECSE_0383 - ECSE_0386	None
eamA	None	ECSE_1623 - ECSE_1628	None
flhE	*yecT*	ECSE_2112	None
cobU	*insH* - *yeeW*	None	ECSF_1852 - ECSF_1896
yibL	None	ECSE_3883 - ECSE_3884	ECSF_3436 - ECSF_3438
ilvN	None	None	ECSF_3517
yjfY	None	ECSE_4497	None
cpdB	*ytfI*	None	None
tsr	None	ECSE_4620 - ECSE_4630	None
ytfK	*ytfI*	None	None
ybfA	*rhsC* - *ybfD*	ECSE_0761 - ECSE_0765	None
ymgE	*ycgY*	ECSE_1244	ECSF_1141 - ECSF_1146
yfgJ	None	None	ECSF_2351 - ECSF_2353
agaI	*yraH* - *yraK*	ECSE_3428 - ECSE_3431	None
potE	*ybfP* - *ybfH*	None	None

3.2.5 Tree split compatibility による樹形比較

本章では、H-NS 結合オーソログ遺伝子配列と H-NS 非結合オーソログ遺伝子配列の大腸菌 44 株における進化速度を比較する。この計算の前提となるのは、どちらの遺伝子群も大腸菌系統群の共通祖先においてすでにゲノム中に存在しており、それらは大腸菌が様々な株に分岐していく過程で垂直的な進化を遂げてきたという仮定である。もしいずれかの遺伝子群でこの仮定が成り立たない場合、それらの遺伝子群の間で配列の多様性を比較したとしても、そのような多様性の差異が進化速度の差異によって生じたという解釈はできなくなる。オーソログ遺伝子を選別する段階で、大腸菌系統群の多数の株に保存されていること、可動性遺伝因子ではないこと、遺伝子重複が生じていないこと、などの厳しい基準によってオーソログ遺伝子を定義したが、依然として、以下に挙げるような進化パターンをたどった遺伝子が含まれている可能性がある。

1. 大腸菌の分岐後に様々な株で独立に獲得された結果、大腸菌系統群の多くの株で保持されるようになった

2. ある大腸菌が外来性遺伝子を獲得し、その遺伝子が大腸菌系統群の中で二次的に伝播していった

以後、このような現象を大腸菌種内 HGT と呼ぶ。大腸菌種内 HGT によって多数の株に保持されるようになった遺伝子の配列を用いて推定される進化系統は、大腸菌の本来の進化系統を反映しない。なぜなら、進化的に離れた大腸菌株間で配列の類似性が非常に高くなりうるためである。したがって、もしもどちらかの遺伝子群で大腸菌種内 HGT が頻繁に生じていたならば、それらの遺伝子群間での多様性の差異は進化速度の差異を意味しない。とりわけ、H-NS 結合遺伝子は外来性遺伝子であることが多いため、H-NS 結合オーソログ遺伝子群において、大腸菌種内 HGT が生じてしまっている蓋然性は否定できない。本研究では遺伝子配列の多様性の比較から進化速度の差異を議論するため、本章で定義したオーソログ遺伝子が H-NS 結合の有無によらず、大腸菌系統群共通祖先の遺伝子であり、かつ、垂直的に受け継がれてきた遺伝子であることを確認することがきわめて重要である。

もしオーソログ遺伝子がすべて大腸菌系統群で垂直的に進化してきたものであるならば、それぞれの遺伝子配列によって推定される系統樹はすべて、同じ樹形となるはずである。遺伝子にかかる選択圧によってそれぞれの進化速度が異なる

としても、その結果は推定された系統樹の枝長の差として観察され、樹形には影響しない。したがって、H-NS 結合オーソログ遺伝子群のそれぞれで推定された大腸菌系統群の系統樹と、H-NS 非結合オーソログ遺伝子群のそれぞれで推定された系統樹の樹形を比較し、それらが一致していれば、H-NS 結合オーソログ遺伝子群において大腸菌種内 HGT が生じていたことを否定できる。だが、大腸菌系統群は進化的にきわめて近縁であるため、それぞれの遺伝子の長さや多重置換の影響により、多くの場合、垂直的な進化をしてきた遺伝子同士であっても系統樹の樹形は完全には一致しない。そこで本研究では、樹形の一致よりも緩和された条件として、Split compatibility を評価する (Nelson-Sathi et al. 2014, Ku et al. 2015)。

Split とは、系統樹に含まれる系統群の二分割のことである（以下、Split の具体例は Holder (2015) による）。系統樹に含まれる枝のそれぞれが、ひとつの Split を表現している。あらゆる樹形の系統樹は Split の集合として表現ができ、逆に Split の集合が情報として与えられると、系統樹の樹形が再構成できる。例として、1 から 6 の数字でラベルされた 6 つの種からなる系統樹を考える。ありうる Split のひとつは、たとえば $\{1,6\}|\{2,3,4,5\}$ である。このとき、系統樹は無根系統樹であることを前提としている。したがって、同一の Split について $\{2,3,4,5\}|\{1,6\}$ という表現も可能である。ここで、この Split を A とおき、Split のそれぞれのグループに含まれるラベルの集合を、$A_1 = \{1,6\}$、$A_2 = \{2,3,4,5\}$ のように定義する。

異なる 2 つの Split は、もしもそれらが同一の系統樹に情報として含まれているならば、“compatible”であると定義される。厳密には、異なる 2 つの Split、A と B があるとき、それぞれの Split で定義されるラベルの集合、A_1, A_2, B_1, B_2 に関して、以下の 4 つの論理積を考え、

$$A_1 \cap B_1 \tag{3.1}$$

$$A_1 \cap B_2 \tag{3.2}$$

$$A_2 \cap B_1 \tag{3.3}$$

$$A_2 \cap B_2 \tag{3.4}$$

上記 4 つの論理積のうち少なくともひとつが空集合 $\emptyset$ であるならば、Split A と B は compatible である。

たとえば、3 つの Split、 A $= \{1,6\}|\{2,3,4,5\}$、B $= \{2,3,4,6\}|\{1,5\}$、C $=$

$\{1,4,5,6\}|\{2,3\}$ について考える。A と B は compatible ではない。なぜなら、

$$A_1 \cap B_1 = \{6\} \tag{3.5}$$

$$A_1 \cap B_2 = \{1\} \tag{3.6}$$

$$A_2 \cap B_1 = \{2,3,4\} \tag{3.7}$$

$$A_2 \cap B_2 = \{5\} \tag{3.8}$$

となるためである。一方、C と A は compatible であり ($A_1 \cap C_2 = \{1,6\} \cap \{2,3\} = \emptyset$)、C と B も compatible である ($B_2 \cap C_2 = \{1,5\} \cap \{2,3\} = \emptyset$)。

以上のように、Split compatibility を計算することで、ある系統樹で観察される二分岐が別の系統樹においても観察されるか否かを検証することができる。これは、ブートストラップ値として知られる、系統樹の各節点における支持度の計算手法のひとつでもある (Yang 2006)。本研究では、垂直的進化の検証手法として近年提案された Split compatibility を用いた計算手法に倣い (Nelson-Sathi et al. 2014, Ku et al. 2015)、ある系統樹のすべての Split のうちの何%が、別の系統樹のいずれかの Split と compatible であるかを計算する、Tree split compatibility を系統樹の一致性のスコアとして使用した。

Tree split compatibility によって H-NS 結合オーソログ遺伝子群の系統樹と、H-NS 非結合オーソログ遺伝子群の系統樹を比較し、それらの樹形が有意に異なるかどうかを検証した。H-NS 結合オーソログ遺伝子は 519 個検出され、H-NS 非結合オーソログ遺伝子は 2,183 個検出されたため、グループを構成する系統樹の数に偏りがある。サンプルサイズバイアスに起因する問題を回避するため、本研究では先行研究 (Nelson-Sathi et al. 2014) に倣い、以下の 5 つの系統樹セットを構築した。

1. Set A: H-NS 非結合オーソログ遺伝子の系統樹（N = 2,183）
2. Set B: H-NS 結合オーソログ遺伝子の系統樹（N = 519）
3. Set C: Set A の系統樹のダウンサンプリング系統樹（N = 519, 非復元抽出）
4. Set D: Set B の系統樹のそれぞれについて HGT イベントのシミュレーション（ランダムに選択した枝の、ランダムに選択した枝への“接ぎ木”）を導入した系統樹（N = 519）

5. Set E: Set B の系統樹のそれぞれについて 2 回の HGT シミュレーションを導入した系統樹（N = 519）

Set A はリファレンスとなる系統樹セットである。Set B は H-NS 結合オーソログ遺伝子の系統樹セット、Set C は H-NS 非結合オーソログ遺伝子の系統樹セットであり、どちらも同数の系統樹から構成される。Set B の Set A に対する Tree split compatibility スコアの分布を、Set C の Set A に対するスコアの分布と比較し、それらの分布が十分に近いものであるかどうかを検証する。Set C は Set A のサブセットであるため、Set C はいわば、Set B の垂直性進化を検証するためのポジティブコントロールとして用いられる。一方、Set D および Set E は、大腸菌種間系統樹に大腸菌種内 HGT をそれぞれ 1 回および 2 回、強制的に導入した系統樹のセットであり、垂直性進化検証のためのネガティブコントロールとして用いられる。したがって、ここで検証するのは、Set B（H-NS 結合）の系統樹が Set C（H-NS 非結合）の系統樹と同程度に近く、わずかにでも HGT が生じた Set D, Set E の系統樹とはまったく異なっている、という結果が観察されるか否かである。この観察が成り立つのであれば、H-NS 結合オーソログ遺伝子群において特に大腸菌種内 HGT が頻繁に生じているという仮説は否定できる。Tree split compatibility スコアの分布の比較は、二群のスコア分布が同一のスコア分布からのサンプルであるという仮定を帰無仮説として、Kolmogorov-Smirnov 検定によって検討した。

3.2.6 大腸菌オーソログ遺伝子のプロテオバクテリア門における保存性

本章で決定した大腸菌オーソログ遺伝子は、少なくとも大腸菌系統群の分岐以降は垂直的に進化してきた遺伝子群であると推測されるが、それよりもさらに以前の HGT によって獲得された遺伝子である可能性はある。そこで、対象とする系統群の範囲を大腸菌 44 株からさらに広げ、プロテオバクテリア門に属する細菌について、大腸菌から進化系統的に離れるに従って、大腸菌オーソログ遺伝子の保存性がどのように減衰していくのかを調べた。本節ではプロテオバクテリア門に属する 48 の細菌株のゲノム情報から、MBGD (the Microbial Genome Database for Comparative Analysis) (Uchiyama 2003, 2007, Uchiyama et al. 2010, 2015) をデフォルトのパラメータで利用してオーソログ遺伝子テーブルを構築した。対象とした 48 の細菌株のうち、ひとつは大腸菌 K-12 MG1655 株、25 株は大腸菌と同じ「科」に属するが異なる「属」に属する細菌（腸内細菌科 *Enterobacteriaceae*）、

14 株は大腸菌と同じ「綱」に属するが異なる「科」に属する細菌（ガンマプロテオバクテリア綱 *Gamma proteobacteria*）、8 株は大腸菌と同じ「門」に属するが異なる「綱」に属する細菌（プロテオバクテリア門 *Proteobacteria*）である。使用した細菌株を表 3.3 に示す。

表 3.3: 遺伝子保存性解析に使用した細菌株

Strains (Taxonomic code in MBGD)
Same species with Escherichia coli
Escherichia coli K-12 MG1655(eco)
Same family but different genus (family Enterobacteriaceae)
Buchnera aphidicola APS (buc)
Citrobacter koseri ATCC BAA-895 (cko)
Enterobacter sakazakii ATCC BAA-894 (esa)
Dickeya zeae Ech1591 (dze)
Edwardsiella ictaluri 93-146 (eic)
Enterobacter sp. 638 (ent)
Erwinia tasmaniensis Et1/99 (eta)
Klebsiella pneumoniae ATCC 700721; MGH 78578 (kpn)
Kosakonia sacchari SP1 (gm03453)
Morganella morganii subsp. morganii KT (mmk)
Pantoea sp. At-9b (pao)
Erwinia carotovora SCRI1043 (eca)
Photorhabdus luminescens TTO1 (plu)
Proteus mirabilis HI4320 (pmr)
Providencia stuartii MRSN 2154 (psi)
Rahnella sp. Y9602 (rah)
Raoultella ornithinolytica B6 ror)
Salmonella typhimurium LT2; SGSC 1412; ATCC 700720 (stm)
Serratia proteamaculans 568 (spe)
Shigella flexneri 301 (sfl)
Escherichia blattae DSM 4481 (ebt)
Sodalis glossinidius morsitans (sgl)
Wigglesworthia glossinidia (wbr)
Xenorhabdus nematophila ATCC 19061 (xne)
Yersinia pestis CO92 (ype)

table continued on next page

continued from previous page

Strains (Taxonomic code in MBGD)
Same class but different family (class Gammaproteobacteria)
Acidithiobacillus ferrooxidans ATCC 53993 (afe)
Aeromonas hydrophila subsp. hydrophila ATCC 7966 (aha)
Shewanella oneidensis MR-1 (son)
Dichelobacter nodosus VCS1703A (dno)
Allochromatium vinosum DSM 180 (alv)
Legionella pneumophila Philadelphia 1 (lpn)
Methylococcus capsulatus Bath (mca)
Carsonella ruddii PV (crp)
Gilliamella apicola wkB1 (gap)
Haemophilus influenzae Rd KW20 (hin)
Pseudomonas aeruginosa PAO1 (pae)
Francisella tularensis Schu 4 (ftu)
Vibrio cholerae N16961 (vch)
Xanthomonas axonopodis 306 (xac)
Same phylum but different class (phylum Proteobacteria)
Acetobacter pasteurianus IFO 3283-01 (apt)
Rickettsia prowazekii Madrid E (rpr)
Burkholderia pseudomallei K96243 (bps)
Neisseria meningitidis MC58 (nme)
Desulfovibrio vulgarish Hildenborough (dvu)
Geobacter sulfurreducens PCA (gsu)
Campylobacter jejuni NCTC 11168 (cje)
Helicobacter pylori 26695 (hpy)

3.2.7 大腸菌相同遺伝子間領域の決定

転写制御領域の大腸菌系統群内多様性を調べるために、本研究では特に、遺伝子間領域（2 つの遺伝子にはさまれた Non-coding 領域）に着目する。そのため、大腸菌系統群の共通祖先が有していた遺伝子間領域（これを以後、相同遺伝子間領域と呼ぶ）を特定したい。しかし、以下に挙げる理由により、相同遺伝子間領域の特定はオーソログ遺伝子の特定以上に注意して行わなければならない。

1. 遺伝子間領域内で相同組換えが生じた可能性がある
2. 配列の挿入・欠失が頻繁に生じている可能性がある
3. 単に遺伝子のアノテーションが適切になされていないために Non-coding とみなされている可能性がある

1 の問題は、遺伝的に不均質なそれぞれ別個の領域によってひとつの領域が構成されてしまうことで、多様性を過大に評価してしまうことにある。2 について、遺伝子間領域はコーディング領域と比較して配列の挿入・欠失が生じやすいと考えられるが、ある領域で挿入・欠失の頻度が高い場合、ギャップ位置前後のアラインメントが難しいという問題により (Lunter et al. 2008)、その領域の多様性をやはり過大に評価してしまう可能性がある。また、3 の場合は、遺伝子間領域の進化を観察していたつもりが、コーディング領域の進化を見ていたことになってしまうため、このような可能性がある領域は避ける必要がある。以上の理由から、本研究では相同遺伝子間領域の採択基準として、以下の 3 つを採用する。

1. 前後に存在する 2 つの遺伝子が共に、前節で定義した、大腸菌全 44 株で保存されているオーソログ遺伝子であり、かつオーソログ遺伝子の並び方もすべての株で等しい
2. その領域を共有している大腸菌株の配列セットの中で、最短の配列と最長の配列の長さが 10%以上異ならない
3. 配列の長さがすべての株で 10bp 以上、300bp 以下である

それぞれの基準の番号が上で述べた問題点に対応している。1 の基準より、前後がオーソログ遺伝子であることから、その間の遺伝子間領域は染色体の相同組換えの影響からは免れていると考えることができる。2 と 3 は一見似ている基準であるが、それぞれが主張していることは異なる。2 は、いずれか 1 株でも、DNA 配列の挿入・欠失によって領域が大きく変化してしまっている場合は、進化距離計算が妥当でなくなる可能性があるため、その領域は相同遺伝子間領域に含めないことを意味する。3 は、進化距離推定に不適当なほど短すぎる領域、そして、アノテーションされていないコーディング領域が含まれ得るような長い領域を取り除くための基準である。大腸菌の遺伝子の平均長はおよそ 1,000bp であるので、上

記の範囲は妥当であると考えられる。

これらの基準によって、712 個の相同遺伝子間領域が抽出された。

3.2.8 相同遺伝子間領域の分子進化解析

相同遺伝子間領域のそれぞれについて、オーソログ遺伝子の進化解析と同様、MAFFT（G-INS-i アルゴリズム） (Katoh and Toh 2008) によるマルチプルアラインメントと、PAML パッケージ (Yang 1997) の Baseml プログラムによる進化距離推定を行った（Baseml のパラメータ設定は以下：tree=ML tree from PhyML, model=7, clock=0, kappa=2.5(starting value), fix_kappa=0(ML estimation of kappa), alpha=0, fix_alpha=1(fixed value), rho=0, fix_rho=1(fixed value), npark=0, nhome=0, Mgene=0.）。遺伝子間領域の進化モデルは GTR（Generalized Time-Reversible） (Tavaré 1986) を用いた。進化距離計算の結果、9 個の領域で、少なくとも 1 つの株ペアの進化距離が 1.0 を超える値となってしまった。これらの領域は、何らかの問題で不適切な配列が含まれてしまっているか、あまりに多様化しているために精確なアラインメントができていない可能性のある領域である。したがってこれらの領域は相同遺伝子間領域セットから取り除き、最終的に、703 個の相同遺伝子間領域セットを構築した。

相同遺伝子間領域のそれぞれについて、“ H-NS が結合している ” か否かの基準は、オーソログ遺伝子の場合と同様、いずれか 1 株で H-NS 結合が見られた場合に、H-NS 結合相同遺伝子間領域であるとした。

遺伝子間領域にかかる選択圧は、その領域内にプロモータ配列をはじめとした遺伝子転写制御領域が存在するか否かによって異なると考えられる。そこで、大腸菌 K-12 株におけるこれらの相同遺伝子間領域内のプロモータ配列の情報を、Regulon DB (Salgado et al. 2013) より取得した。遺伝子間領域内のプロモータ配列の有無に基づいて、それらの間の相同遺伝子間領域の多様性を比較した。

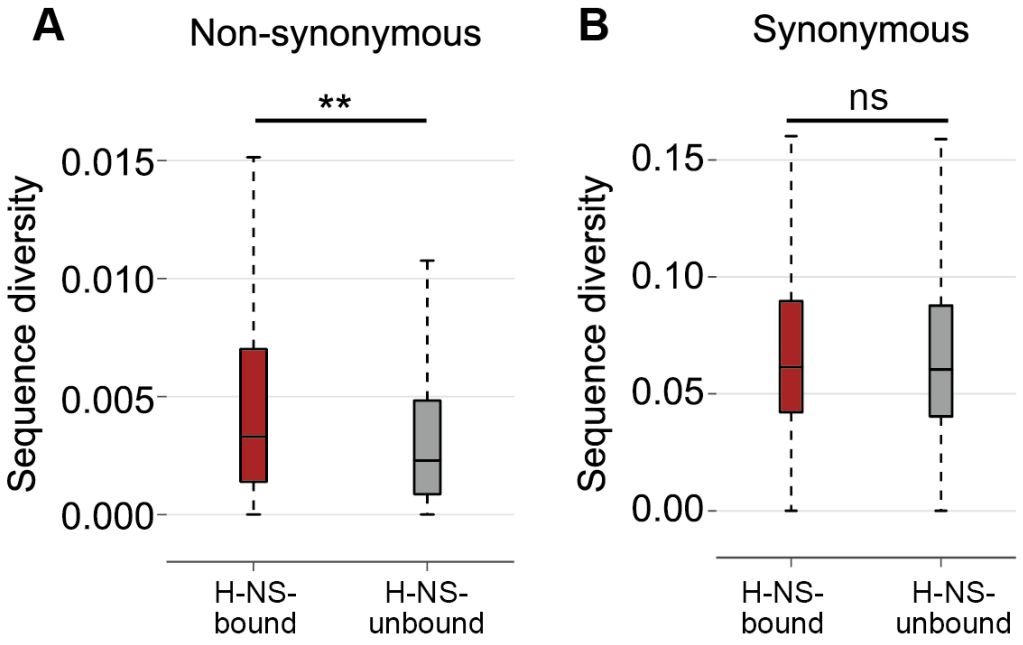

図 3.5: H-NS 結合オーソログ遺伝子（赤色、N=519）と H-NS 非結合オーソログ遺伝子（灰色、N=2,183）の (A) 非同義座位および (B) 同義座位における多様性 $^{**}p\text{-}value < 0.001$, ns; not significant (Wilcoxon rank sum test)

3.3 結果

3.3.1 H-NS 結合オーソログ遺伝子の多様化

大腸菌 44 株のオーソログ遺伝子を、「H-NS 結合」と「H-NS 非結合」に分類し、非同義座位（Non-synonymous）多様性と同義座位（Synonymous）多様性を比較した結果を、図 3.5 に示す。

図 3.5A と図 3.5B の縦軸のスケールは 10 倍異なり、非同義座位と比較して同義座位の多様性は大きいことがわかる。これは、コーディング領域において一般的に見られる特徴であり (Sharp 1991)、変異が即、指定するアミノ酸の変化を引き起こす非同義座位に対して、同義座位では変異が指定するアミノ酸の変化を伴わないため、変異の蓄積速度が速い傾向にある。

大腸菌 44 株のオーソログ遺伝子では、同義座位において、H-NS 結合、非結合

の多様性の分布に統計的に有意な差は見られなかった（図 3.5B）。一方、非同義座位においては、H-NS 結合オーソログ遺伝子は、H-NS 非結合オーソログ遺伝子と比較して、配列多様性が有意に増大していた（$p\text{-}value < 0.001$, ウィルコクソン順位和検定）。

同義座位の多様性にほとんど違いが見られず、非同義座位において H-NS 結合オーソログ遺伝子に選択的に変異が蓄積していることから推測されるのは、H-NS 結合領域に見られる多様性の増大は、H-NS の結合による物理的な影響によって生じているのではないという可能性である。もし H-NS が、結合している DNA 配列に物理的に変異を導入するような役割を果たしているのであれば、同義座位、非同義座位といった、コードされている情報に依存した分布の差は生じえない。つまり、多様性の増大は遺伝子機能にかかる選択圧の緩和によって生じているのかもしれない。これは、H-NS による転写抑制が、その遺伝子の進化に影響をおよぼしている可能性を示唆している。

3.3.2 進化解析の妥当性検証 1：H-NS 結合の定義

ここで、“H-NS が結合している”とした際の定義に問題があったかもしれない。ここでの定義は、H-NS が遺伝子のどの場所に結合しているかは斟酌せず、ORF 内を 1bp でも H-NS が覆っていたら「H-NS 結合」としていた。H-NS がどのようなメカニズムで転写を抑制するのか、完全に明らかになっているわけではないが、H-NS がプロモータ領域周辺に結合して RNA ポリメラーゼの結合を物理的に阻害するか (Oshima et al. 2006)、H-NS が DNA 結合ドメインと多量体形成ドメインを持つために生じる DNA-H-NS-DNA ブリッジの形成によって、プロモータ領域周辺に DNA のループ構造を形成し、RNA ポリメラーゼをその中にトラップしてしまう (Dame et al. 2005, 2006) モデルが提案されている。H-NS の DNA への結合定数はそれほど大きくないことが知られており、遺伝子の 3’ 側にわずかに結合している程度ならば、RNA ポリメラーゼが H-NS を解離させながら進行し、転写を問題なく完遂するかもしれない。もしそのような結合パターンの遺伝子が、上で定義した“H-NS 結合オーソログ遺伝子”に含まれていたならば、非同義座位の多様性を H-NS による転写抑制に関連づける上記のような議論は適切ではない可能性がある。

そこで、H-NS が ORF の 3' 側に結合し、かつ、その遺伝子を覆っている H-NS 結合領域が ORF の長さの 10%以下であった場合は、“H-NS 非結合遺伝子”として分類し直した場合の多様性の分布を計算した（図 3.6A）。結果は、図 3.5 と変わらず、そのような可能性を考慮しても依然として、非同義座位の多様性に有意な差が見られた。

また、細菌は、複数の遺伝子がひとつのプロモータによって制御され、ひとつの転写単位として同時に転写される、オペロンとして知られるシステムを持っている。そのため、転写抑制の影響を議論するならば、H-NS が直接結合している遺伝子だけでなく、その遺伝子自体に H-NS が結合していなくとも、その遺伝子の発現に影響を及ぼすプロモータ領域に H-NS が結合しているならば、“H-NS 結合遺伝子”とみなすべきだという考え方もあり得る。そこで、大腸菌の転写単位およびプロモータ領域の情報を RegulonDB (Salgado et al. 2013) より取得し、H-NS がプロモータ領域に結合していた場合は、そのプロモータによって制御される転写単位に含まれるすべての遺伝子を“H-NS 結合遺伝子”として分類し、H-NS がオペロンの途中に結合していた場合は、そのオペロンに含まれ、かつ、H-NS 結合よりも下流に位置する遺伝子を“H-NS 結合遺伝子”として分類して、多様性の分布をそれ以外の H-NS 非結合遺伝子と比較した（図 3.6B）。この場合も、“H-NS 結合遺伝子”における非同義座位の多様性の増大が見られた。一方で、同義座位に関してはこれまでと結果が異なり、非同義座位ほどの顕著な差ではないものの、統計的に有意に H-NS 結合遺伝子において多様性が増大していた。同義座位における多様性の増大に関しては考察で議論する。

以上より、H-NS 結合オーソログ遺伝子における非同義座位多様性の増大は、“H-NS 結合遺伝子”の定義をより厳しく、あるいは、定義を緩和しても観察された。

3.3.3 進化解析の妥当性検証 2：アラインメントの信頼性

配列アラインメント中にギャップが含まれると、その周辺でのアラインメントの精度が下がり、結果として多様性を過大に見積もってしまう危険性が知られている (Lunter et al. 2008)。H-NS 結合オーソログ遺伝子で非同義座位多様性が増大していたのは、単に H-NS 結合オーソログ遺伝子のアラインメントにおいてギャップが多く含まれていたためであるかもしれない。そこで、オーソログ遺伝子

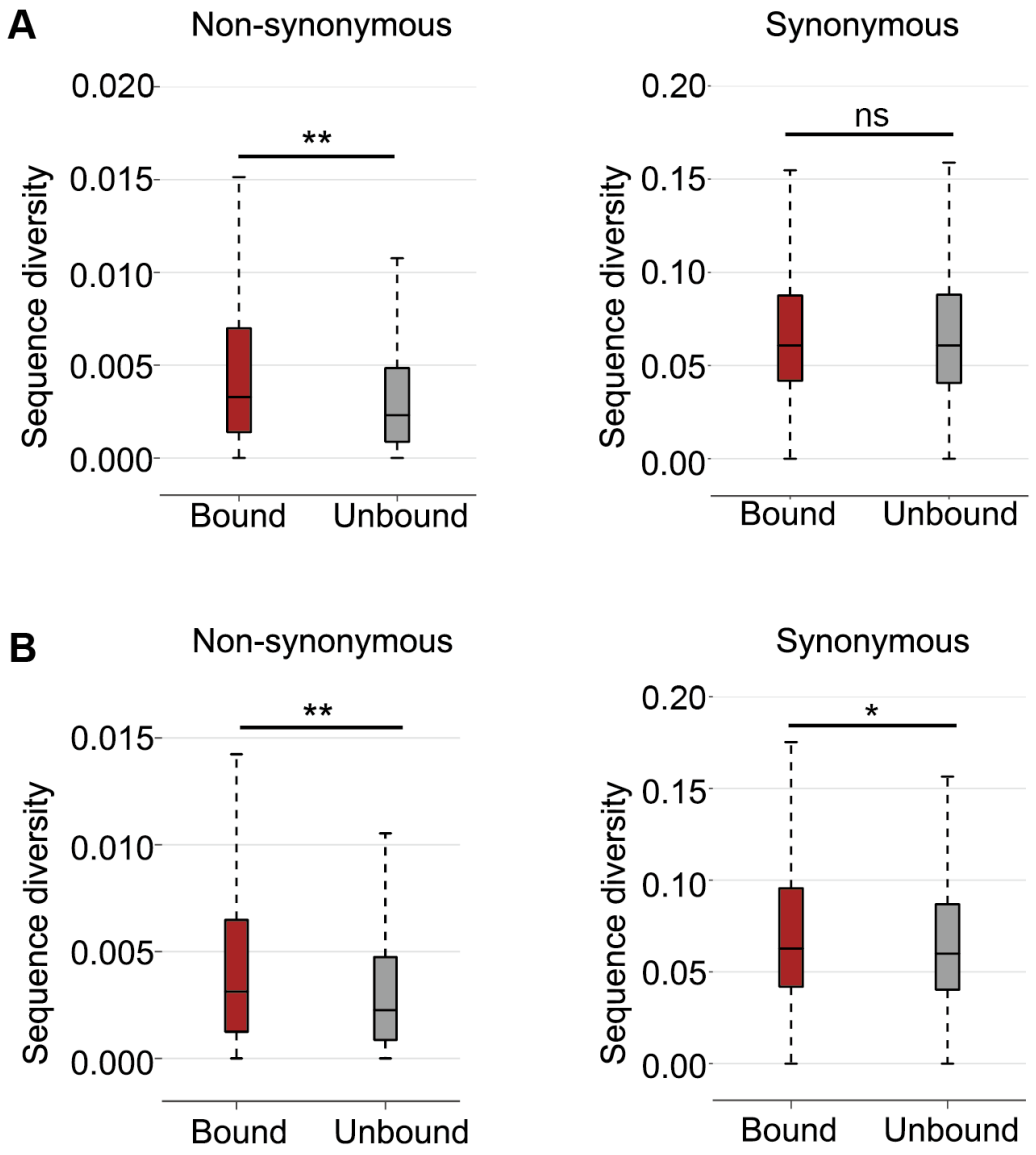

図 3.6: (A)3' 側 10%以内の H-NS 結合を無視した場合の、H-NS 結合オーソログ遺伝子（赤色, N=474）と H-NS 非結合オーソログ遺伝子（灰色, N=2,228）の多様性 (B) 転写単位を考慮した場合の、H-NS 結合オーソログ遺伝子（赤色, N=752）と H-NS 非結合オーソログ遺伝子（灰色, N=1,950）の多様性 $^{**}p\text{-}value < 0.001$, ns; not significant (Wilcoxon rank sum test)

のアラインメントをすべて検証し、その中にギャップが含まれていないオーソログ遺伝子のみを抽出して、同様に H-NS 結合の有無における多様性の比較を行った (図 3.7)。アラインメントにギャップが含まれないオーソログ遺伝子のみで比較し

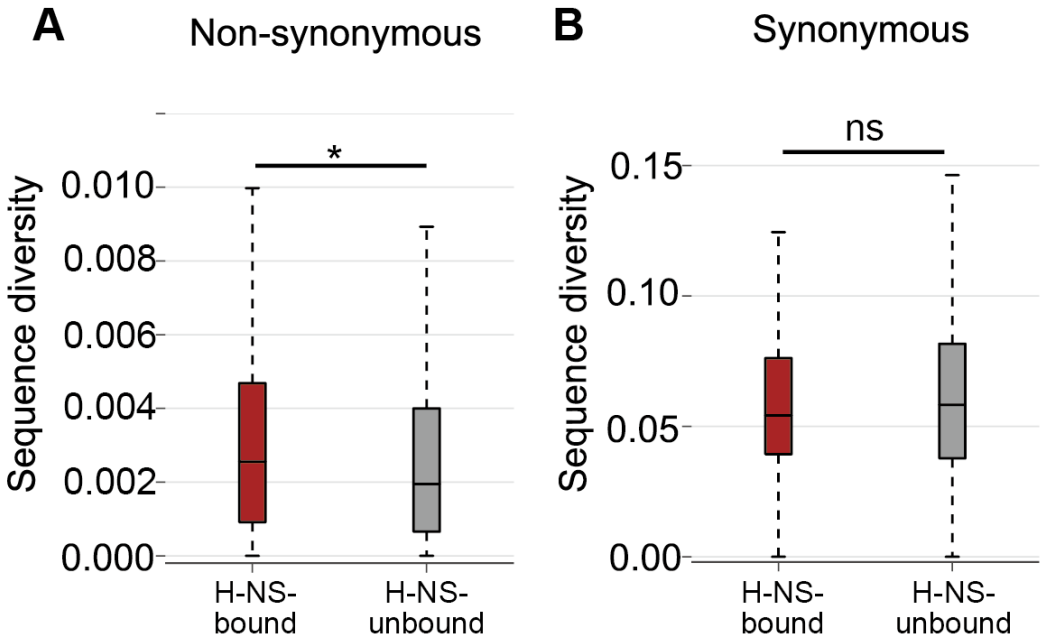

図 3.7: ギャップなしアラインメントのみのオーソログ遺伝子の多様性　H-NS 結合オーソログ遺伝子（赤色, N=159）と H-NS 非結合オーソログ遺伝子（灰色, N=940）の (A) 非同義座位および (B) 同義座位における多様性 *p-$value$ < 0.05, ns; not significant (Wilcoxon rank sum test)

ても、H-NS 結合遺伝子では非同義座位の多様性が H-NS 非結合遺伝子と比べて有意に増大していた。したがって、H-NS 結合遺伝子における多様性増大は、ミスアラインメントによって生じたものではないことがわかった。

3.3.4 進化解析の妥当性検証 3：H-NS 結合遺伝子の大腸菌種内 HGT の可能性

H-NS 結合オーソログ遺伝子で観察される多様性の増大を、進化速度の増大と解釈するためには、H-NS 結合・非結合のオーソログ遺伝子群がどちらも大腸菌系統群内で垂直的に進化してきたものであることを確認しなければならない。H-NS 結合遺伝子は外来性遺伝子であることが多く、それぞれの株で独立に獲得された遺伝子であったり、大腸菌系統群内での組み換えの頻度が高い遺伝子の可能性があるため、この点は特に注意を要する。H-NS 結合・非結合に関わらず、どちらの遺

伝子群も大腸菌系統群の共通祖先においてすでに存在していた遺伝子であることを確認するため、それぞれの遺伝子群に含まれる各遺伝子で推定された系統樹の樹形の安定性を Tree split compatibility で検討した。リファレンスとなる H-NS 非結合オーソログ遺伝子との Tree split compatibility スコアの比較によって、H-NS 結合オーソログ遺伝子において大腸菌種内 HGT が頻繁に生じているかを検証できる。結果を図 3.8 に示す。

両側 Kolmogorov-Smirnov 検定の結果、H-NS 非結合オーソログ遺伝子の系統樹セットと、H-NS 結合オーソログ遺伝子の系統樹セットで、それらのスコアが同一の分布からサンプリングされたものであるという帰無仮説は棄却できなかった（$p\text{-}value = 0.16$）。一方で、ほんのわずかでも系統樹に変更（一回の HGT シミュレーション）を加えると、帰無仮説は強く棄却された（$p\text{-}value < 0.001$）。したがって、H-NS 非結合オーソログ遺伝子と比較して H-NS 結合オーソログ遺伝子で大腸菌種内 HGT が頻繁に生じているとは言えない、という結果が得られた。

スコアの分布は、H-NS の結合の有無に関わらず、小さい値に集中している。これは、スコアとして Minimal tree split compatibility (Nelson-Sathi et al. 2014) を使用しているためである。Minimal tree split compatibility は、系統樹のすべての split の中で、リファレンス系統樹セットと一致する割合が最も小さいものを、その系統樹のスコアとする手法である。したがって、系統樹の一致性のスコアとしては厳しい基準となる。また、大腸菌系統群はお互いに進化的に近縁であるため、多くの遺伝子で系統樹に進化的な情報がほとんど含まれない（大腸菌系統群を再現するのに十分な塩基置換が遺伝子内に生じていない）ことがスコアに影響している可能性がある。とはいえ、一本の枝の付け替えというわずかな操作（HGT シミュレーション）によって系統樹の一致性は大きく損なわれることから、H-NS 結合オーソログ遺伝子と H-NS 非結合オーソログ遺伝子で系統樹の樹形の分布に大きな差はないといえる。

以上より、オーソログ遺伝子のそれぞれの系統樹の樹形がすべて一致しているとは言えないものの、少なくとも H-NS の結合の有無で樹形に大きな偏りは見られず、H-NS 結合オーソログ遺伝子で大腸菌種内 HGT が頻繁に生じているとはいえない、という結論が得られた。

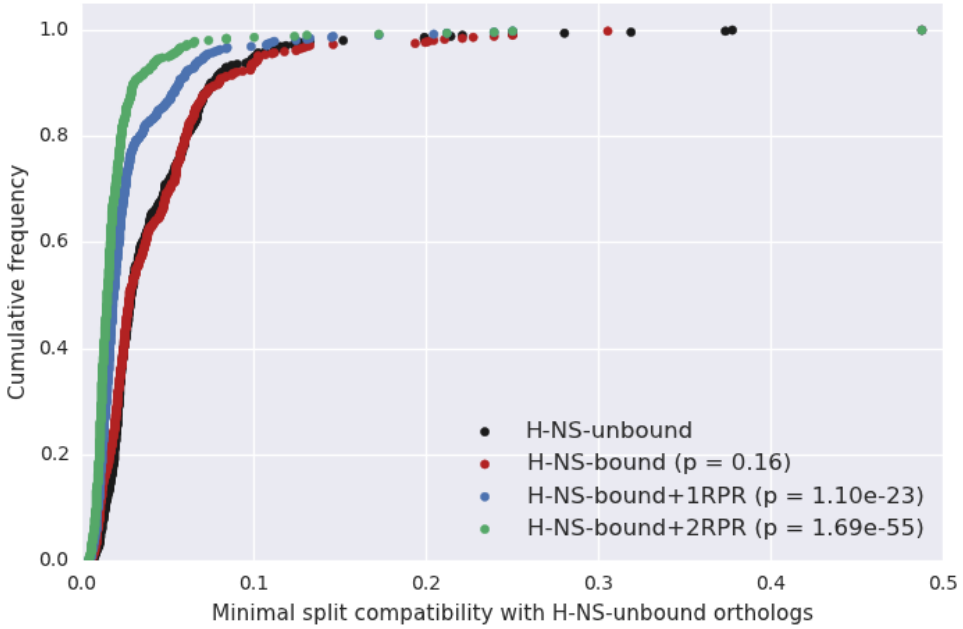

図 3.8: Minimal tree split compatibility の累積分布　黒点：Set C(H-NS 非結合) のスコアの分布　赤点：Set B(H-NS 結合) のスコアの分布　青点：Set D(H-NS 結合オーソログ遺伝子に 1 回の HGT シミュレーションを加えた系統樹) のスコアの分布　緑点：Set E(2 回の HGT シミュレーションを加えた系統樹)　p-value は Set C(H-NS 非結合) との両側 Kolmogorov-Smirnov 検定による

3.3.5　進化解析の妥当性検証 4：外来性遺伝子の特徴によるものか？

H-NS は外来性遺伝子に結合しやすいことが知られているため、このような H-NS 結合オーソログ遺伝子に見られる進化の加速は、単に外来性遺伝子の特徴を反映しているだけかもしれない。先行研究によって、*Bacillus cereus* (Hao and Golding 2006) および大腸菌系統群 (Davids and Zhang 2008) において、近年獲得された外来性遺伝子が急速に変異を蓄積している傾向が確認されている。先行研究で報告されているこれらの遺伝子においてもこれまでの議論と同様に H-NS や、H-NS に似た「外来性遺伝子を抑制する」機能を持つタンパク質によって抑制されているために進化が加速している可能性はあるが、H-NS とは独立に、外来性遺伝子の特徴としてこのような進化が生じている可能性は否定できない。外来性遺伝子がゲノムへ組み込まれたあと、それ自体として急速に変異を蓄積させる特徴を持つならば、H-NS による抑制は必ずしも外来性遺伝子の進化に必要ではないかも

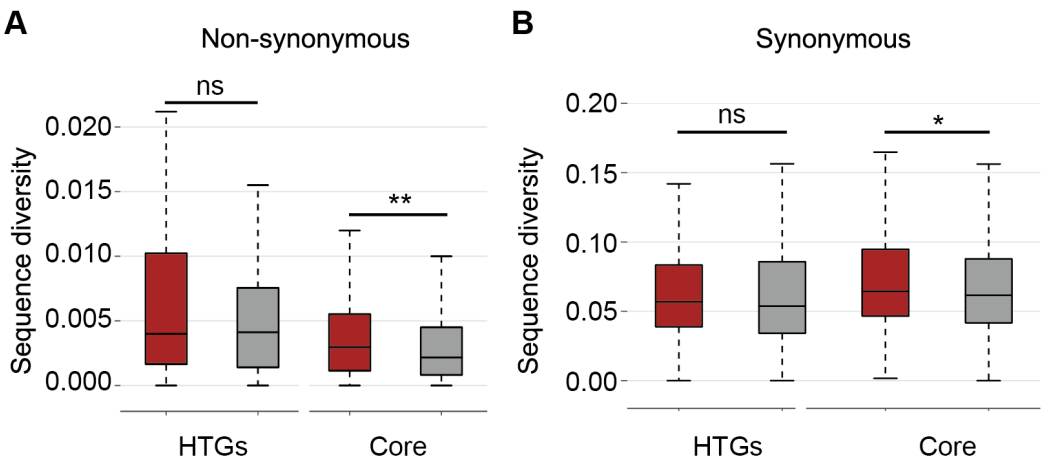

図 3.9: 外来性オーソログ（HTGs, 赤色：H-NS 結合 N=230, 灰色：H-NS 非結合 N=271）とコアオーソログ（Core, 赤色：H-NS 結合 N=289, 灰色：H-NS 非結合 N=1,912）の非同義座位多様性（A）および同義座位多様性（B）**p-$value < 0.001$, *p-$value < 0.05$, ns; not significant(Wilcoxon rank sum test)

しれない。外来性遺伝子は宿主のゲノムとは異なった塩基組成やコドン使用頻度である場合が多いため、それを宿主のゲノムに適応させる Amelioration と呼ばれる進化プロセスなどがその背景のメカニズムとなりうる。

そこで本節ではこれまでに扱ってきたオーソログ遺伝子を「外来性オーソログ」と「コアオーソログ」に分け、それぞれで、これまでに見たような H-NS の結合による多様性の差が観察できるかどうかを検討する。外来性オーソログは、過去の 3 つの先行研究 (Lawrence and Ochman 1998, Garcia-Vallve et al. 2003, Nakamura et al. 2004) で「外来性遺伝子」であると見なされた大腸菌遺伝子であり、コアオーソログはそれ以外の大腸菌遺伝子とした。これらの 3 つの先行研究では、塩基組成やコドン使用頻度の偏りから大腸菌系統群内の外来性遺伝子を特定している。本研究では外来性とコアを大腸菌オーソログ遺伝子について定義するため、外来性とはいっても、少なくとも大腸菌系統群が分岐する以前に獲得されたものであると仮定していることには注意が必要である。

外来性オーソログとコアオーソログの非同義座位多様性、同義座位多様性を比較したのが図 3.9 である。　まず、一般的な傾向として、外来性オーソログ（HTGs）の非同義座位多様性は H-NS 結合の有無に関わらずコアオーソログ（Core）と比較して大きかった (図 3.9A、HTGs と Core の比較)。これは先行研究 (Hao and

Golding 2006, Davids and Zhang 2008) によって報告された外来性遺伝子の多様性増大傾向と一致している。本研究ではさらに、コアオーソログにおいて、H-NS が結合している場合に多様性が統計的に有意に増大している傾向が観察された（図 3.9A $p\text{-}value < 0.001$、Core の中で H-NS 結合（赤色）と H-NS 非結合（灰色）の比較）。コアオーソログにおけるこのような傾向は、外来性遺伝子が急速に変異を蓄積するという仮説のみによっては説明できない。また、コアオーソログについては同義座位においても、H-NS の結合による多様性の差が観察された（図 3.9B、Core）。したがって、H-NS 結合オーソログ遺伝子における多様性の増大は単に近年獲得された外来性遺伝子の進化の傾向のみによって生じているのではなく、コアオーソログについても、非同義座位、同義座位のそれぞれで H-NS の結合に依存した進化の加速が生じていることがわかった。

一方、外来性オーソログの中では、H-NS の結合の有無による多様性の統計的な差は見いだされなかった（図 3.9A, B の HTGs）。しかし、集団の分散の等質性を検定するルビーン検定によって、外来性オーソログに関して H-NS 結合の有無で非同義座位多様性の分散に差があるか否かを検定した結果、統計的に有意に H-NS 結合外来性オーソログで分散が大きいことがわかった（図 3.9A, ルビーン検定 $p\text{-}value = 0.010$, 赤色（H-NS 結合）の箱の上側のエッジ＝上側四分位数が、灰色（H-NS 非結合）の上側四分位数と比較して大きい）。したがって、統計的に有意な差があるとまでは言えないものの、外来性オーソログにおいても、H-NS が結合することによってある程度の割合の遺伝子の多様性が増大していることがわかった。

コアオーソログにおいて H-NS 結合による進化の加速が観察されたが、コアオーソログはいったいどのような遺伝子なのだろうか？H-NS が基本的に外来性遺伝子に結合すること、さらに、第 2 章の結果より、H-NS の結合が長期的に維持されることから考えると、おそらく H-NS が結合しているコアオーソログは、外来性オーソログよりもさらに古い時代に大腸菌の祖先系統で獲得された、「古代の外来性遺伝子」であると考えられる。この仮説を検討するために、進化解析の対象とする細菌群の系統範囲をさらに広げ、大腸菌が属するプロテオバクテリア門において、上記 4 つのカテゴリの大腸菌オーソログ遺伝子（H-NS 結合 HTGs, H-NS 非結合 HTGs, H-NS 結合 Core, H-NS 非結合 Core）の保存性が大腸菌からの進化距離が遠くなるにしたがってどのように減衰していくのかを比較した。対象としたのは、

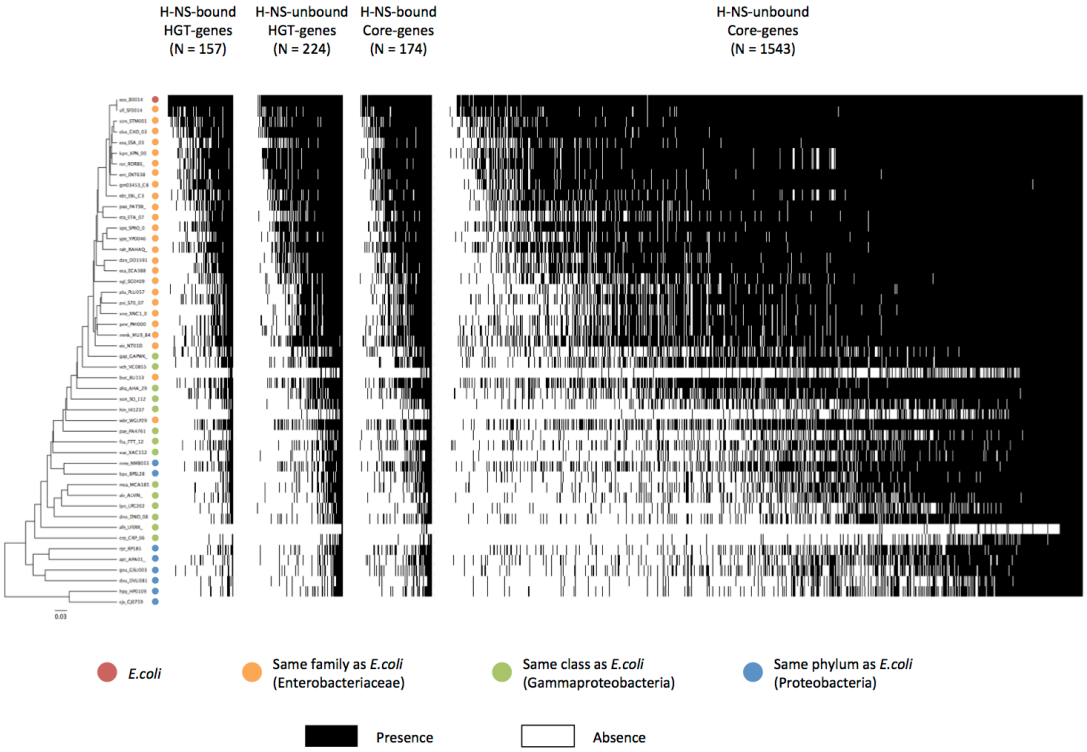

図 3.10: プロテオバクテリア門における大腸菌オーソログの保存性　各列がそれぞれ 4 つのカテゴリの大腸菌オーソログ遺伝子、各行が解析に使用したプロテオバクテリア門の株を表している。黒色は、そのオーソログ遺伝子がその株のゲノムに存在していることを示し、白色はゲノム上に存在していないことを示している。左端のデンドログラムは、DnaK タンパク質配列を用いて推定したプロテオバクテリア門細菌の進化系統樹である

大腸菌 K-12 MG1655 株、大腸菌と同じ「科」に属するが異なる「属」に属する細菌 25 株（腸内細菌科 *Enterobacteriaceae*）、大腸菌と同じ「綱」に属するが異なる「科」に属する細菌 14 株（ガンマプロテオバクテリア綱 *Gamma proteobacteria*）、大腸菌と同じ「門」に属するが異なる「綱」に属する細菌 8 株（プロテオバクテリア門 *Proteobacteria*）の合計 48 株である（表 3.3）。MBGD を利用してすべての株のオーソログ遺伝子を決定し、本章で決定した大腸菌オーソログ遺伝子 2,702 個と一致する MBGD オーソログを抽出して、4 つのカテゴリに属する遺伝子群をそれぞれの株が持っているのか、あるいは欠失しているのかを調べた（図 3.10）。
対応するオーソログ遺伝子の、各株のゲノム上における有無を観察すると、4 つの

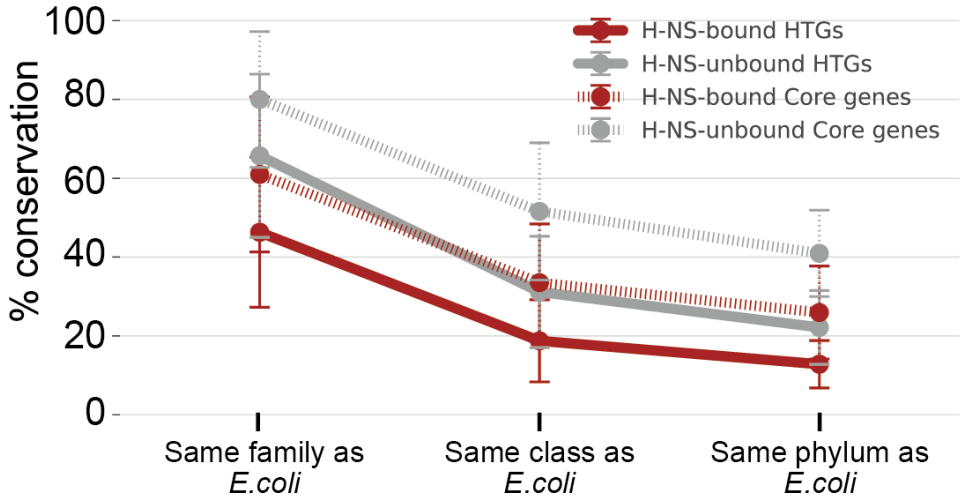

図 3.11: 大腸菌オーソログ保存性の減衰　大腸菌オーソログの保存率を、Family *Enterobacteriaceae*、Class *Gamma proteobacteria*、Phylum *Proteobacteria* のそれぞれで平均化した。赤色実線：H-NS 結合外来性オーソログ、灰色実線：H-NS 非結合外来性オーソログ、赤色点線：H-NS 結合コアオーソログ、灰色点線：H-NS 非結合コアオーソログ

カテゴリ（H-NS 結合 HTGs、H-NS 非結合 HTGs, H-NS 結合 Core, H-NS 非結合 Core）間で、保存性の減衰の度合いが異なるように見える。

そこで、4 つのカテゴリのオーソログ遺伝子について、同じ「科」だが異なる「属」の系統群、同じ「綱」だが異なる「科」の系統群、そして同じ「門」だが異なる「綱」の系統群のそれぞれにおける平均保存率を比較した（図 3.11）。　保存率の減衰は、H-NS 結合外来性オーソログ（図 3.11 赤色実線）が最も速かった。これは、H-NS 結合外来性オーソログが比較的近年に大腸菌系統群の共通祖先で獲得された遺伝子であることを示唆している。また、それとは逆に H-NS 非結合コアオーソログ（図 3.11 灰色点線）は保存率の減衰が遅く、プロテオバクテリア門の広い細菌で保存されていた。この遺伝子群には必須遺伝子をはじめとして細菌に広く保存されている遺伝子が多く含まれるので、これらは大腸菌系統群共通祖先のずっと以前から維持されてきた遺伝子であるのかもしれない。興味深いことに、H-NS 結合コアオーソログ（赤色点線）はそれらの中間的な保存率であり、H-NS

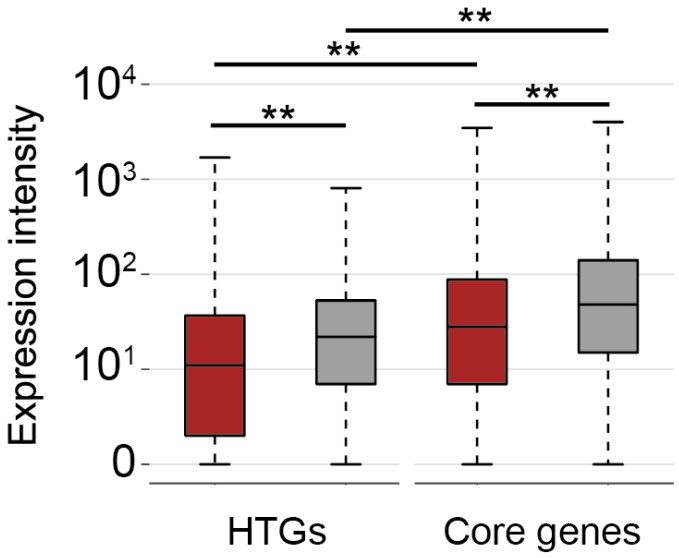

図 3.12: 先行研究 (Taniguchi et al. 2010) の RNA-seq に基づいた 4 種類の大腸菌オーソログ遺伝子の発現量分布の比較　HTGs: 外来性オーソログ、Core:コアオーソログ、赤色:H-NS 結合遺伝子、灰色:H-NS 非結合遺伝子 $^{**}p\text{-}value < 0.001$, wilcoxon rank sum test

結合外来性オーソログよりは減衰が遅いが、H-NS 非結合コアオーソログよりか減衰が速かった。したがって、H-NS 結合コアオーソログは、H-NS 結合外来性オーソログよりも古い時代に獲得された「古代の外来性遺伝子」である可能性が高い。

H-NS は外来性遺伝子の転写を抑制する。それは、外来性遺伝子がゲノムに取り込まれた初期段階において、その発現が宿主にとって有害である場合が多いためである。さらに本章の解析で、H-NS が結合した遺伝子は、近年の水平伝播で獲得された遺伝子だけでなく、はるか以前に獲得された遺伝子においても、多様性が増大している傾向が示された。このような変異の蓄積が適応的な意義を持っているか否かは、それらの遺伝子の転写量を比較することによって評価することができる。4 つのカテゴリの遺伝子群について、RNA-seq データ (Taniguchi et al. 2010) をもとに大腸菌 K-12 株における発現強度を比較したのが図 3.12 である。

発現量の比較によって、H-NS 結合の有無に関わらず、コアオーソログは外来性オーソログと比較して統計的に有意に発現量が高いことがわかった（図 3.12 $p\text{-}value < 0.001$）。すなわち、H-NS 結合コアオーソログは、外来性オーソログと

比較してより宿主のシステムに適応していると考えられる。一方で、おそらく H-NS が結合している影響によって、H-NS 結合コアオーソログは H-NS 非結合コアオーソログよりも発現量は低かった（図 3.12 Core genes の比較　$p\text{-}value < 0.001$）。4 つのカテゴリの遺伝子群の発現量は、4 つのカテゴリの非同義座位多様性の傾向と反比例している。すなわち発現量が低い遺伝子ほど、非同義座位の多様性が高い傾向が見られる。

本節の解析結果をまとめる。まず、H-NS 結合遺伝子では H-NS 非結合遺伝子と比較して、コアオーソログにおいては統計的に有意に、外来性オーソログにおいてはわずかに、非同義座位多様性が増大している傾向が観察された。H-NS 結合外来性オーソログは比較的近年の遺伝子水平伝播によって、H-NS 結合コアオーソログはそれよりも以前の遺伝子水平伝播によって獲得された遺伝子であることが推定された。H-NS 結合コアオーソログでは外来性オーソログと比較して発現量が上昇していた。以上の結果は、H-NS 結合コアオーソログがはるか以前に大腸菌系統群共通祖先において獲得され、H-NS が結合して発現を抑制することによって徐々に変異を蓄積し、少しずつ宿主のシステムに統合されて発現量が上がっていった、というプロセスによって説明できる。H-NS の結合が長期的に維持されていることと考え合わせると、おそらく H-NS 結合コアオーソログにおいて進化が加速している要因は、H-NS による転写抑制の効果で有害な変異による悪影響を回避し、進化の過程でこれらの遺伝子がより“自由に”変異を蓄積して宿主のゲノムに適応しているためだと考えられる。

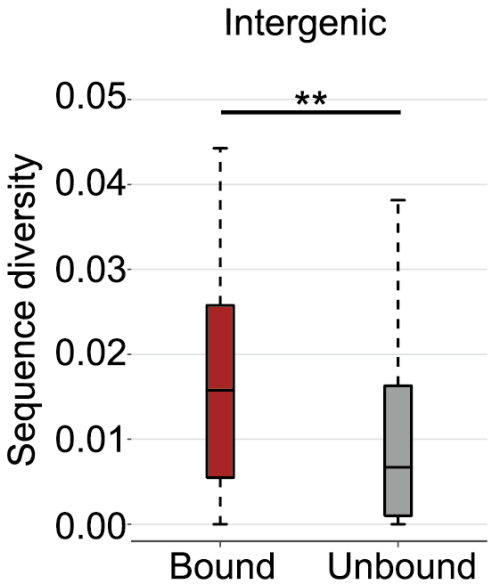

図 3.13: H-NS 結合相同遺伝子間領域（赤色, N=94）と H-NS 非結合相同遺伝子間領域（灰色, N=609）の多様性 $^{**}p\text{-}value < 0.001$ (Wilcoxon rank sum test)

3.3.6 H-NS 結合相同遺伝子間領域の多様化

次に、大腸菌 44 株の相同遺伝子間領域を「H-NS 結合」と「H-NS 非結合」に分類し、多様性を比較した結果を図 3.13 に示す。

ここでも、オーソログ遺伝子の非同義座位の場合と同様の観察が得られ、H-NS が結合している相同遺伝子間領域の多様性は、H-NS が結合しない相同遺伝子間領域の多様性と比較して、有意に増大しており、およそ 2 倍の値を示していた。

相同遺伝子間領域においては特に、配列の挿入・欠失によって、アラインメント内にギャップが生じやすくなると考えられるため、相同遺伝子間領域に関してもアラインメント中にギャップを含まない相同遺伝子間領域のみで、H-NS 結合の有無による多様性の差を検定した（図 3.14）。アラインメントにギャップを含まない相同遺伝子間領域のみを用いても H-NS が結合している領域における多様性の増大が観察された。したがって、領域中に生じる配列の挿入・欠失のみの影響でこのような差が生じているのではなく、実際に領域内において塩基置換頻度が増大

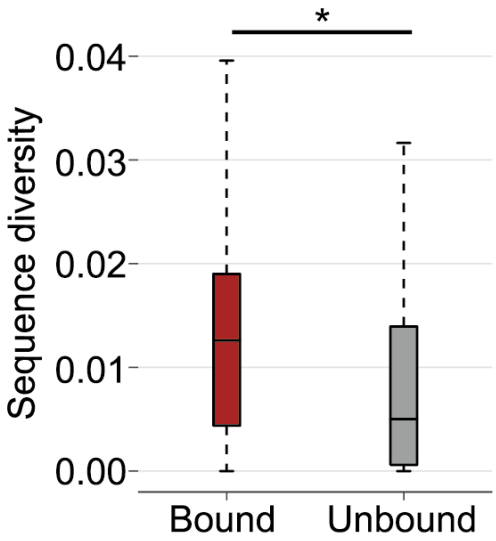

図 3.14: アラインメント中にギャップを含まない相同遺伝子間領域を抽出した場合の H-NS 結合相同遺伝子間領域（赤色, N=56）と H-NS 非結合相同遺伝子間領域（灰色, N=458）の多様性 *p-$value$ < 0.05 (Wilcoxon rank sum test)

していることが確認できた。

オーソログ遺伝子の場合と同様、この結果を遺伝子の転写調節と関連させて議論したい。つまり、H-NS 結合相同遺伝子間領域では、H-NS 非結合の領域と比較して、転写調節に関わるような領域の配列が多様化している一方で、そうでない領域における変異の蓄積速度は H-NS の有無によらず一定である、といった観察を得ることが可能だろうか。残念ながら、1bp ごとの情報の冗長性を評価できるコーディング領域と異なり、遺伝子間領域（Non-coding 領域）ではどの位置がどれほどの情報を担っているのかはそれほど明確ではない。プロモータ領域に関して言えば、σ 因子が結合する配列など、制御に重要な配列要素のいくつかは明らかになっているが、多様な転写調節因子の結合配列のすべてが明らかになっているわけではない。そのため、遺伝子間領域の中で、どの配列が制御の多様化に必須であり、どの配列はある程度の多様性が許容されうるのかを分類して評価するのは困難である。

そこで本研究では、より簡単な分類方法として、「転写制御配列が存在し得るか

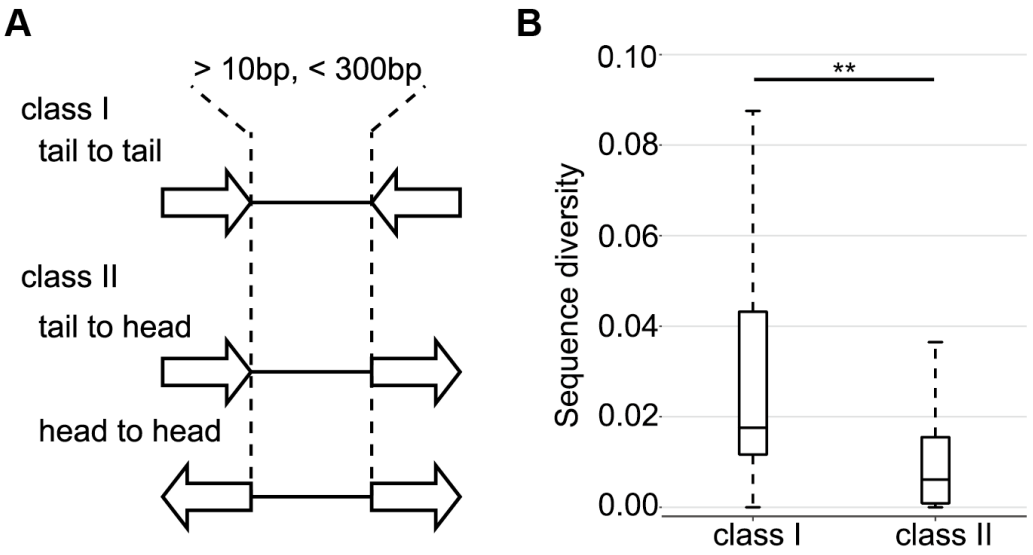

図 3.15: (A) 相同遺伝子間領域の分類。挟まれている 2 つの遺伝子の転写方向によって、class I, class II の 2 つに領域を分類した (B)class I 相同遺伝子間領域（左, N=92）と class II 相同遺伝子間領域（右, N=611）の多様性 $^{**}p\text{-}value < 0.001$ (Wilcoxon rank sum test)

否か」を軸として遺伝子間領域を分類する（図 3.15A）。

遺伝子間領域は、その前後に位置する 2 つの遺伝子の転写方向によって、tail to tail, tail to head, head to head の 3 パターンに分類できる（図 3.15A）。head to head は 2 つの遺伝子の転写開始点が共に含まれているような領域である。この領域は、制御に必要な情報の密度が最も高い（領域中の配列のうち、制御配列の占める割合が大きい）と考えられる。tail to head は、プロモータ配列が存在する場合と、存在しない場合の 2 つの可能性が考えられる。前後の遺伝子が異なる転写単位として制御されるのであれば、プロモータ配列を含む制御領域はこの領域に存在することになる。この領域と前後の遺伝子が共にひとつのオペロンとして転写されるのであれば、この転写単位の制御領域はもっと上流に存在する可能性が考えられる。特徴的だと思われるのは、tail to tail の遺伝子間領域である。この領域は、双方の遺伝子にとっての転写終結点であり、転写開始点はほとんどの場合存在し得ない。近年、Pervasive transcription として、通常の機能遺伝子のプロモータ

配列とは別に、Open reading frame 内、あるいは遺伝子の転写終結点にプロモータ配列が存在し、アンチセンス RNA が転写されて相補的な遺伝子の mRNA と結合することで転写制御に密接に関わっている例がいくつか知られているが (Wade and Grainger 2014)、それがどれほど普遍的な現象であるかは明らかになっておらず、基本的には tail to tail の領域内には転写制御領域はほとんど存在しないと思われる。したがって、すべての相同遺伝子間領域を、tail to head, head to head が含まれる class II と、tail to tail からなる class I に分類すると、前者はプロモータ配列などの制御領域が存在し得る遺伝子間領域であり、後者は転写調節に関わる制御領域がほぼ存在しない遺伝子間領域として分類することができる。

この分類に基づいて、それぞれの相同遺伝子間領域の多様性を比較した結果、class II では大腸菌株間の多様性が低く抑えられている一方、class II と比較して class I は有意に多様化していることがわかった（図 3.15B）。class I 遺伝子間領域は、制御領域を含まないが、2 つの遺伝子の転写を終結させるターミネータ配列は存在し得る領域である。ターミネータ配列は、ρ 因子依存性のターミネータであるか、内在性のターミネータであるかによっても異なり、コンセンサス配列が明らかにされているわけではないものの、GC-rich である、あるいは T-tract（T 塩基の連続性）が必要であると考えられている (Peters et al. 2011)。だが、ターミネータ配列は、厳密な制御が必要とされる遺伝子の転写開始点と比べて、とにかく RNA ポリメラーゼの進行を止めれば良いので、ある程度の配列多様性は許容されるのではないかと思われる。そのため、図 3.15B のように、制御領域が含まれ得る class II 遺伝子間領域では多様性が低く、その一方、class I 遺伝子間領域では多様性が高い傾向が見られたのだと考えられる。

以上より、class I と class II の分類を用いることによって、オーソログ遺伝子の場合の「変異が遺伝子機能に直接影響を及ぼし得る非同義座位」「変異が必ずしも遺伝子機能に影響を与えない同義座位」という対応関係と類似した関係性が、相同遺伝子間領域の場合、「変異が転写制御に直接影響を及ぼし得る class II 遺伝子間領域」「変異が必ずしも転写制御に影響を与えない class I 遺伝子間領域」という対応関係として得られる。

この分類に基づいて、それぞれの class で、H-NS が結合している相同遺伝子間領域と H-NS 非結合の相同遺伝子間領域にさらに分類し、多様性の分布を比較してみる。その結果が、図 3.16 である。

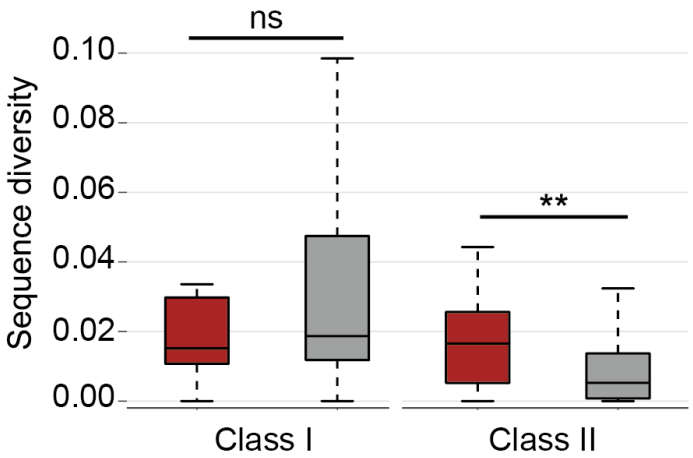

図 3.16: class I 相同遺伝子間領域（赤色：H-NS 結合 N=14, 灰色：H-NS 非結合 N=78）と class II 相同遺伝子間領域（赤色：H-NS 結合 N=80, 灰色：H-NS 非結合 N=531）の配列多様性の分布 $^{**}p\text{-}value < 0.001$, ns; not significant (Wilcoxon rank sum test)

class I 相同遺伝子間領域においては、H-NS の結合の有無によって、多様性の分布に差は見られなかった（図 3.16）。一方で、class II 相同遺伝子間領域においては、H-NS が結合している場合、多様性は有意に増大していた（$p\text{-}value < 0.001$, ウィルコクソン順位和検定）。これは、オーソログ遺伝子の解析で議論した解釈と、ほぼ同様の解釈ができる結果である。すなわち、制御領域の存在しない class I 遺伝子間領域では、H-NS の結合があってもなくても、多様性の分布に差が見られないことから、H-NS の結合が直接に変異の導入に関わっているわけではない。一方で、様々な制御配列が存在する可能性のある class II では H-NS の結合によって多様性が増大している。これは、H-NS の結合によって生じたその遺伝子の発現抑制の影響によって、その領域に存在する制御配列、たとえば転写調節因子の結合位置、プロモータ配列、attenuation site、pausing site などの配列の多様化が許容されているという解釈が可能である。

相同遺伝子間領域においてもオーソログ遺伝子の議論と同様、HGT によって前後の遺伝子とともに取り込まれた遺伝子間領域が、Amelioration などの作用で進化速度が速く見えてしまっている可能性を検討する必要がある。H-NS 結合遺伝子

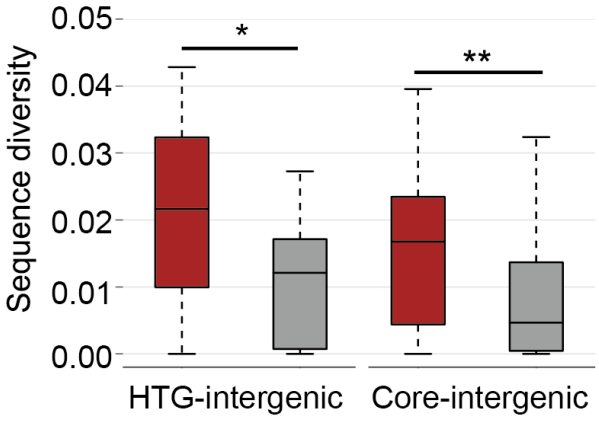

図 3.17: 外来性遺伝子間領域（HTG-intergenic, 赤色：H-NS 結合 N=24, 灰色：H-NS 非結合 N=16）とコア遺伝子間領域（Core-intergenic, 赤色：H-NS 結合 N=28, 灰色：H-NS 非結合 N=411）の多様性 $^{**}p\text{-}value < 0.001$, $^{*}p\text{-}value < 0.05$ (Wilcoxon rank sum test)

間領域における多様性の増大が、その領域が比較的近年の HGT で獲得された外来性配列であることに起因しているのか、それとも H-NS が結合することで選択圧が緩和されて変異が蓄積しているのかを区別するため、ここでは本章で定義した相同遺伝子間領域を、「外来性遺伝子間領域」と「コア遺伝子間領域」に分類し、それぞれの領域における H-NS の作用を比較する。なお、転写制御領域に蓄積する多様性を観察するために、以下の解析では class II 相同遺伝子間領域のみを扱う。外来性遺伝子間領域は、先行研究 (Lawrence and Ochman 1998, Garcia-Vallve et al. 2003, Nakamura et al. 2004) と本章の解析によって特定された「外来性オーソログ遺伝子」に挟まれている領域である。コア遺伝子間領域は、その前後に位置するオーソログ遺伝子がどちらもコアオーソログ遺伝子である領域である。前後のオーソログ遺伝子の一方が外来性オーソログ遺伝子であり、もう一方がコアオーソログ遺伝子であるような遺伝子間領域は、その領域内で組み換えが生じた可能性があるため、解析対象から除外した。外来性遺伝子間領域とコア遺伝子間領域の配列多様性比較の結果を図 3.17 に示す。

H-NS 結合外来性遺伝子間領域における配列多様性は、H-NS 非結合外来性遺

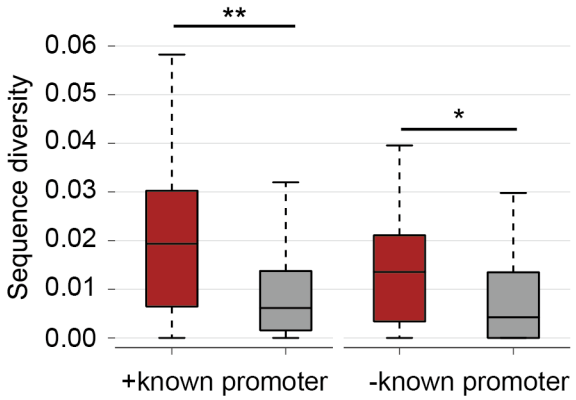

図 3.18: プロモータ配列の有無による class II 遺伝子間領域の多様性比較　+known promoter:　プロモータ配列あり（赤色：H-NS 結合 N=50, 灰色：H-NS 非結合 N=267）、-known promoter:プロモータ配列なし（赤色：H-NS 結合 N=30, 灰色：H-NS 非結合 N=264）　**p-$value$ < 0.001, *p-$value$ < 0.05 (Wilcoxon rank sum test)

伝子間領域の配列多様性と比較して有意に増大していた（図 3.17 HTG-intergenic, p-$value$ < 0.05）。さらに、コア遺伝子間領域においても、H-NS が結合することで有意に配列多様性が増大している傾向が観察された（図 3.17 Core-intergenic, p-$value$ < 0.001）。したがって、相同遺伝子間領域における H-NS が結合している場合の進化の加速は、HGT によって獲得された配列が急速にゲノムに適応していく過程のみによっては説明ができないことがわかった。遺伝子間領域においてはたしかに、H-NS 結合領域に限定して進化が加速していた。

class II 遺伝子間領域には、転写制御領域を含み転写開始点として機能する場合と、オペロンの途中に位置して転写制御領域を含まない場合の 2 通りがありうる。H-NS によって多様化している領域がすべて後者であったならば、そのような多様性獲得の適応的意義は小さい。H-NS の結合が転写制御領域の存在する遺伝子間領域において生じていることを確認するために、RegulonDB (Salgado et al. 2013) から取得した大腸菌 K-12 株のプロモータ配列位置情報を参照し、class II 相同遺伝子間領域を、プロモータ配列あり、およびプロモータ配列なし、の 2 種類に分類した。その 2 つのカテゴリのそれぞれで、H-NS 結合の有無による配列多様性を

比較した。結果を図 3.18 に示す。

解析の結果、H-NS 結合の有無による配列多様性の差異は、むしろプロモータ配列が存在する遺伝子間領域でさらに大きくなっていることがわかった。したがって、多様性の増大はジャンク DNA においてのみ生じているのではなく、H-NS が結合している場合には、制御領域が存在する遺伝子間領域でかえって進化が加速している傾向が観察された。

3.4 考察

本章では、大腸菌系統群のオーソログ遺伝子と相同遺伝子間領域における分子進化解析から、大腸菌の様々な株が分岐する以前、共通祖先が有していたと思われる遺伝子および転写調節領域において、H-NS が結合している場合、多様性が大きい傾向にあることを明らかにした。同義座位の多様性の分布は H-NS 結合の有無によって大きな差は見られず、H-NS 結合の定義として転写単位を考慮した場合、および H-NS 結合コアオーソログに関して、多様性がわずかに高かった。一方、非同義座位における多様性は、H-NS 結合オーソログ遺伝子で大きく増加していた。相同遺伝子間領域においては、転写制御配列が存在しない遺伝子間領域（tail to tail）の多様性の分布は H-NS 結合の有無によって差が見られず、制御配列が存在する（head to head)、あるいはオペロンとして前後の遺伝子と共に転写される可能性のある遺伝子間領域（tail to head）では H-NS が結合している場合に多様性が増加していた。これらの結果からわかるのは、オーソログ遺伝子、相同遺伝子間領域が共に、多様性の分布について共通のパターンを示していることである。すなわち、

1. 同義座位、class I 遺伝子間領域など、表現型に与える影響が少ない領域では、H-NS 結合の有無で多様性の分布の差は小さい

2. 非同義座位、転写制御領域など、表現型に直接的に影響を与え得る領域では、H-NS が結合している場合に多様化する傾向がある

の 2 点である。

1 から、H-NS の DNA 配列への結合が、なんらかの物理化学的理由によって直接的に、その結合領域下の DNA 配列に変異を導入する可能性は低い。もし H-NS

が積極的に変異の蓄積を促すのであれば、その結合 DNA 配列のコードする意味内容（同義座位か非同義座位か、ジャンク DNA か制御配列であるか）に関わらず、H-NS 結合配列において多様性が増加しているはずである。実際は、同義座位、class I 遺伝子間領域では多様性にほとんど差が見られなかった。

多様性に差が見られるのは、非同義座位、制御領域など、大腸菌の生存に直接的に影響を与え得る領域だった。非同義座位では、塩基置換が即、指定するアミノ酸の変化を引き起こし、それは結果として、遺伝子がコードしているタンパク質の性質を変化させることに繋がる可能性がある。制御領域における塩基置換は、転写調節因子や RNA ポリメラーゼの DNA への結合のしやすさに変化をもたらし、その結果、遺伝子の発現量が大きく変化する可能性がある。つまり、非同義座位や制御領域における塩基置換は、大腸菌の表現型に与える影響が比較的大きい。H-NS 結合の有無によって多様性の差が生み出されるのは、このような領域である。

通常、遺伝子機能や転写制御などに密接に関わる DNA 配列は、中立的に変異し得る領域と比較して、多様性が低く保たれている。これは、有害突然変異を排除する「純化選択」が働くためである。コーディング領域の非同義座位では、塩基の置換が直接的にコードしているタンパク質の機能の変化を引き起こし、そのような変化は生存にとって有害である可能性が高い。その結果、純化選択の効果により、非同義座位の多様性は低く抑えられる (Sharp 1991)。実際に、非同義座位と同義座位における多様性の分布の平均値には、10 倍もの開きがあった（図 3.5A,B、H-NS 非結合遺伝子の比較）。同様の傾向は遺伝子間領域についても当てはまり、class I 領域（tail to tail）の多様性は、class II 領域（tail to head, head to head）と比較して有意に大きい（図 3.16、H-NS 非結合遺伝子間領域の比較）。これは先行研究でも報告されている傾向であり (Molina and van Nimwegen 2008, Tsoy et al. 2012)、class II 領域の多様性が低いのは、遺伝子の転写制御に関わる配列に生じた純化選択の結果として解釈できる。

ある情報に対する純化選択は、その情報が「評価される」ことがなければ、効果を発揮し得ない。ある個体で、ある遺伝子に変化が生じた場合、その遺伝子が発現し、生育に悪影響を与える、あるいは致死的であってはじめて、純化選択によってその新しい遺伝子は排除される。その新しい遺伝子が発現することがなければ、純化選択の網にかかることはない。進化というプロセスにおいて、実は「自然選択」の効果は限定的であり、必ずしもすべての情報が自然選択の効果を受けるわ

けではない。厳しく自然選択の効果に従うのは、「評価される」情報のみである。これは中立理論の骨子であり (Kimura 1984)、現代進化論のすべての基礎である。

本章の解析の結果、純化選択の効果によって多様性が低く抑えられているはずの領域において、H-NS が結合している場合、多様性が増大する傾向がみられた。これは、H-NS の結合が純化選択の効果を減少させていることを強く示唆している。つまり、H-NS が結合し、遺伝子の転写を抑制することによって、本来「評価されるべき領域」が「評価されない領域＝中立的な領域」のように振る舞い、その結果、その領域にかかるべき純化選択が働かなくなり、多様性が増大していることを意味している。

この仮説は、転写単位を考慮して“H-NS 結合遺伝子”を定義した際の多様性比較の結果によっても支持される（図 3.6B）。転写単位を考慮した H-NS 結合遺伝子の定義では、ある遺伝子のコーディング領域内に H-NS が結合していなくても、その遺伝子を含むオペロンの制御領域、あるいはオペロンの上流側に H-NS が結合していた場合は、H-NS によって転写抑制を受けている遺伝子であると考え、“H-NS 結合遺伝子”として扱っている。そのため、H-NS の結合による直接的・物理的影響を受けていない遺伝子が、転写単位を考慮した“H-NS 結合遺伝子”には多く含まれているが、その場合においても非同義座位多様性に大きな差が観察された (図 3.6A)。したがって、やはり非同義座位の多様性は H-NS の結合によって直接もたらされるのではなく、遺伝子の発現を抑制するという間接的な効果によってもたらされている可能性が高い。

転写活性の高い遺伝子では進化が遅く、一方、転写活性の低い遺伝子が多様化しやすい傾向は以前から知られており (Drummond et al. 2005)、そのような傾向が見られる理由として、以下の仮説が提案されている。

1. タンパク質の機能的制約による純化選択（上記と同様の議論）
2. 転写とカップリングした変異修復機構の存在
3. mRNA の構造や安定性にかかる選択
4. 翻訳効率にかかる選択
5. 翻訳ロバストネスにかかる選択

2、および 3 の仮説は、本章の結果から考えると、たとえ影響があったとしてもその効果はそれほど大きくないと考えられる。転写時に変異を修復するシステムや、mRNA の安定性などは、同義座位、非同義座位といった DNA 配列の意味的な区分とは無関係であり、もし転写活性の高い遺伝子における配列の保存性がそれのみによって規定されているのだとしたら、図 3.5A, B に見られるような、同義座位・非同義座位間の多様性の差が説明できない。

4 の翻訳効率にかかる選択とは、発現の高い遺伝子は同義座位も含めて使われやすいコドンを使っており、それに合わせるために、非同義座位の変異にも制約がかかるという仮説である。たとえ同じアミノ酸を指定している場合でも、それぞれの種によって使われやすいコドンと使われにくいコドンが存在し、そのようなコドン使用頻度の違いは、遺伝子の翻訳効率に影響する (Sharp et al. 1988)。したがって、転写活性の高い遺伝子では翻訳効率を高く維持する選択がかかり、多様性が低く抑えられることが予想される (Akashi 2003)。5 の翻訳ロバストネスとは、Drummond らによって提案され (Drummond et al. 2005)、維持されるのは翻訳効率ではなく、あるアミノ酸が別のアミノ酸に置き換わってしまったとしても、本来のタンパク質と同じ立体構造を取り得るようなロバストな配列が、高発現の遺伝子には望まれる傾向にあるという仮説である。4 および 5 の仮説は、タンパク質の機能に対する純化選択でないが、翻訳効率、翻訳ロバストネスなど、広い意味での遺伝子の性質に対する純化選択である。転写単位を考慮した H-NS 結合遺伝子（図 3.6）、および H-NS 結合コアオーソログ（図 3.9）では、わずかではあるが統計的に有意に、同義座位においても多様性が増加していた。H-NS 結合遺伝子における同義座位多様性の増大は、このような自然選択からの回避によって、生じていたのかもしれない。いずれにしても、4 および 5 によって高発現遺伝子の多様性が低く抑えられているという仮説は、本章で示した結果と矛盾するものではない。重要な点は、H-NS が結合した非同義座位において、このような選択の影響から逃れ、多様性が増大していることであり（図 3.5A）、さらに、H-NS 結合領域の多様性増大が、遺伝子配列のみならず、遺伝子間領域についても同様に観察され、それがやはり、転写制御に意味を持つ配列に限定して生じている点である（図 3.16）。

また本章ではオーソログ遺伝子を、比較的近年に HGT によって獲得された「外来性オーソログ」と、それ以外の「コアオーソログ」に分類し、H-NS 結合コアオーソログは H-NS 非結合コアオーソログと比較して多様性が高いことを明らかにし

た（図 3.9）。さらに、H-NS 結合コアオーソログはおそらく、外来性オーソログよりも以前に HGT によって獲得された「古代の外来性遺伝子」であることが推測された（図 3.11）。一方で、H-NS 結合コアオーソログは、H-NS が結合しているにも関わらず、外来性オーソログと比べて発現量が高い傾向にあった（図 3.12）。H-NS 結合コアオーソログは、H-NS による長期的な転写抑制の結果、多様性が増大し、おそらく H-NS と競合する転写活性化因子の獲得などによって、発現量の最適化に成功したのだと考えられる。

以上は、H-NS による転写抑制を軸とした外来性遺伝子の「純化選択の緩和」と「宿主のシステムへの適応」によって、統一的に解釈されうる。本章の結果は、Dorman の提唱した仮説 (Dorman 2007) に基づいた H-NS の役割を、強く支持するものであるといえる。

第4章　H-NS結合配列のk-mer特徴

本章では、H-NS結合領域下のDNA配列を詳細に解析することによって、第2章で観察したH-NS結合領域およびH-NS結合強度の高度な保存性が、どのような配列要素の保存によって達成されているのかを調べる。多変量解析の結果から、H-NS結合領域の位置および結合強度を維持しつつも、第3章で見られたような配列の多様化が許容されるメカニズムの解明を目指す。

4.1　緒言

第2章の解析の結果、大腸菌3株の“Common”ゲノム領域上で、H-NSの結合が高度に維持され、それらはDNA配列の保存によって達成されていることが示唆された。すなわち、ゲノム上でH-NSが結合する位置と領域の長さを規定する何らかのDNA配列要素が存在し、3株の“Common”領域ではそのような配列が保存されていると考えられる。一方で第3章の解析の結果、H-NS結合領域においては、コーディング領域、遺伝子間領域によらず、大腸菌系統群の中でDNA配列が多様化している傾向が見いだされた。これら2つの観察結果は、一見矛盾しているように感じられる。ゲノム上の同じ領域について、前者はDNA配列の強固な維持を、後者はDNA配列の多様性の増大を意味しているためである。DNA結合タンパク質は通常、ゲノム上の何らかの結合モチーフ（A, T, C, Gの4文字から構成される何らかの文字列、あるいは文字列の組み合わせ）を認識してDNAに結合するが、上記2点の矛盾を解消するためには、H-NSはそのような単純な配列モチーフに基づいて結合するのではない、と結論せざるを得ない。そうではなく、DNAの4文字配列としては多様化を許容しつつも、認識に必要とされる情報は維持されるような、何らかの配列パターンを、H-NSは認識していると考えられる。

H-NSが結合するために必要なDNA配列要素について、これまで多数の研究がなされてきた。H-NSはDNA配列のminor grooveに結合すると考えられており、さらには、intrinsically bent DNA（DNA結合タンパク質などの外的な作用

の影響ではなく、DNA 配列そのものの文字列の並び方の影響で屈曲している二本鎖 DNA）における minor groove に結合しやすいと考えられている (Spurio et al. 1997)。だが、H-NS 結合配列としてコンセンサスの得られている配列モチーフはこれまで存在していない。先行研究では、以下のような配列要素が提案されている。

1. AT 含量（Adenine および Thymine 塩基が含まれる割合）の高い領域 (Lucchini et al. 2006, Navarre et al. 2006)
2. A-tract 配列（AnTm, n+m が 4 以上で TpA を含まない） (Crothers et al. 1990, Zuber et al. 1994, Oshima et al. 2006)
3. ATATA など、AT 含量の高い配列の中に存在する TpA 配列 (Kahramanoglou et al. 2011, Gordon et al. 2011)

1 は、H-NS の結合が、DNA 配列に含まれる A あるいは T 塩基の割合が高い領域に見いだされる傾向が強い、という主張である。だが、「AT 含量」という指標は結合配列要素として明確ではない。分母の取り方（10 bp なのか 1 kbp なのか）によって、つまり、局所的 AT 含量なのか、あるいは、ある程度の広さの領域中の AT 含量なのかによって、「AT 含量」の値は大きく変動し得る。2 および 3 は、同様に A/T に関連した配列であるが、配列の要素により強い制限がある。A-tract 配列は、もともと intrinsically bent DNA の代表的な例として、DNA 二本鎖の立体構造の観点から研究され (Crothers et al. 1990)、後に、ゲノム上の A-tract 配列存在領域が、H-NS 結合領域と良く相関することが見いだされた (Zuber et al. 1994, Oshima et al. 2006)。A-tract は、A が 4bp 以上、あるいは T が 4bp 以上並んだ配列であり、かつ、その途中に TpA ステップ（T の直後に A が続く 2 塩基）を含まない配列である。AT 塩基対の構造、および、DNA が塩基対のスタッキングを揃えることにより構造を安定化させる働きにより、AA（TT）あるいは AT の 2 連続塩基がある程度以上の長さで連続している領域では、DNA 構造が物理的に屈曲する（図 4.1）。

一方、3 の TpA 配列は、2 の A-tract とは対立する仮説である。Kahramanoglou ら (Kahramanoglou et al. 2011) は、H-NS 結合領域に共通して存在する配列モチーフとして、ATATA を報告している。Gordon らによる研究 (Gordon et al. 2011) では合成 DNA 配列断片を用いた *in vitro* 実験により、H-NS に高親和性を持つ DNA 配列は A-tract 配列ではなく、A-tract 配列を含まない AT-rich 配列（TpA ステップ

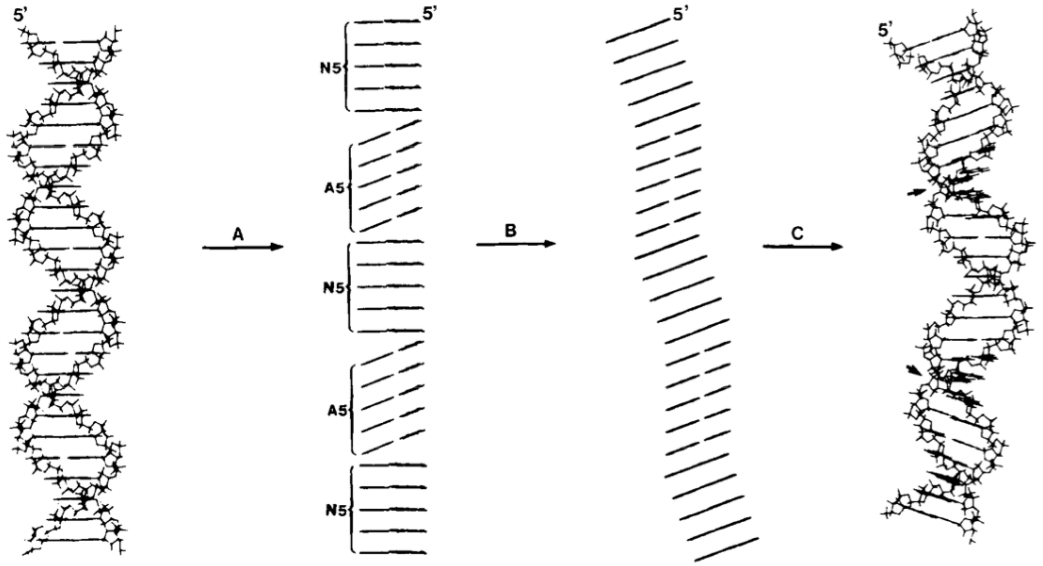

図 4.1: A-tract 配列によって DNA 二本鎖構造が屈曲する理由の模式図。N5 が任意の 5bp、A5 が A-tract の条件を満たす 5bp である。文献 (Crothers et al. 1990) より引用。

が多く存在する）であり、同程度の AT 含量ならば、A-tract の存在はむしろ H-NS の結合にネガティブに働く、と報告している。

いずれにしても現時点では、H-NS が AT 含量の高い DNA 配列に結合しやすいことは確かであるものの、H-NS が実際に何を認識して結合しているのかについてはコンセンサスが得られていない。*in vivo* における H-NS 認識配列の特定が難しい理由として、H-NS が多量体化することによって DNA への結合を安定化させていること、また、おそらく H-NS は高親和性配列を最初に認識して結合し、そこをコアとして H-NS 多量体を形成して周囲に広がっていくこと（「核形成・伝播」モデル (Lang et al. 2007)）などにより、H-NS の結合領域と、高親和性配列の存在位置が必ずしも対応していないことが挙げられる。また、そもそも H-NS の結合領域は（おそらく、外来性遺伝子であることによって）本質的に AT-rich であるため、単純に「H-NS 結合領域」と「H-NS 非結合領域」の DNA 配列を比較し、H-NS 結合領域に頻出する配列を特定する方法（Kahramanoglou ら (Kahramanoglou et al. 2011) はそのようにして ATATA 配列を得ている）では、AT-rich な領域に確率的に現れやすい配列が炙り出されてしまい、H-NS 結合領域「内」の「高親和性配列」と「多量体化することで弱く結合を保っている領域」とを区別することができな

い。

本章では、以上の問題点を解決し、H-NS の高親和性 DNA 配列を特定するために以下の 2 つの戦略をとる。

1. H-NS 結合「強度」情報を積極的に利用し、H-NS 結合領域内の高親和性領域を特定する

2. 1 bp から 6 bp の可能なすべての配列に対して、H-NS 結合強度の回帰モデルを構築し、モデル比較によって高親和性配列を特定する

1 について、ChIP-seq 法の利点は、DNA タンパク質の結合領域と非結合領域との区別ができること以上に、局所的な「結合強度」について定量的なデータが得られる点にある。以下、H-NS「結合強度」という単語を、ChIP-seq 実験の結果得られた共沈降配列断片のゲノム上のエンリッチメントを意味する単語として使用する。結合強度、すなわち ChIP-seq 実験で得られる波形の強度は、化学的な「結合親和性」や「結合定数」とは異なった、独自の情報を有する量である。同じくシーケンス情報をベースとした RNA-seq の場合は、cDNA 配列のリファレンスゲノムへのマッピングの結果得られる波形強度が、それぞれの遺伝子の「発現量」の良い近似値として利用される。同様に ChIP-seq の場合は、DNA 結合タンパク質と共沈降した DNA 断片の、リファレンスゲノムへのマッピングの結果得られる波形強度は、実験に用いた細胞集団の中で、DNA 結合タンパク質が特定のゲノム領域に結合している細胞がどれほどの割合で存在しているのか、という情報を反映している。つまり、全細胞で DNA 結合タンパク質の結合が生じているようなゲノム領域では結合強度は高く、逆に、ある領域で DNA 結合タンパク質が結合していたり、結合していなかったりする、言い換えれば、DNA 結合タンパク質が「剥がれやすい」領域では、結合強度は弱くなるはずである。ChIP-seq 法のこのような特徴は、特に H-NS 結合領域の解析において、有用である。なぜなら、第 2 章で見たように、本質的に H-NS が結合しないような領域ではノイズレベルの結合強度変動が観察されるが、その一方で、H-NS 結合領域では非常に広いスケールの結合強度が観察されたためである（図 2.4）。これらは、ゲノム上の各位置における、H-NS の「剥がれやすさ」を意味する情報だと考えることができる。また、H-NS 結合強度波形に見られるパターンとして、各領域の広さ（H-NS 結合の広がり）は領域ごとに多様であったが、広い結合領域であっても、その全域にわたって特定の結合

強度でべったりと結合しているのではなく、ピーク位置と、その両脇でなだらかに減衰する結合、というパターンがいくつか連続し、その重ね合わせの結果として広い結合領域が形成されているように観察できる点が特徴的である（図 2.7）。つまり、長く広がった H-NS 結合領域についても、その内部の結合ピーク位置を特定し、その位置の DNA 配列を解析することによって、高親和性配列を特定することが可能であるかもしれない。さらに重要な点は、このようなピーク位置のそれぞれの相対的な結合強度が、独立に行った再現性実験の間で、あるいは複数株の実験においてさえ、強く相関していた点である（図 2.3）。これは、相対的な結合強度の値が、実験時のランダムな変動ではなく、何らかの生物学的な特性を反映していることを強く示唆している。以上より、H-NS 結合配列の解析においては、結合強度情報を明示的に取り扱うことが重要であると考える。

2 について、本研究では、そのような結合強度を応答変数として、DNA 配列を説明変数とした回帰モデルを導入する。本研究で仮定していることは、1)H-NS 結合強度の変動は本質的に DNA 配列によって規定されており、2) 結合強度の高い位置では高親和性配列が密集し、低い位置では高親和性配列が少ない、という 2 点である。高親和性配列としては、H-NS が二本鎖 DNA の minor groove に結合することが知られているため、二本鎖 DNA の 1 ピッチ（10.5 bp）よりも短い 1 bp から 6 bp として考える。ATGC 文字列の 1 bp から 6 bp すべての並び方について、「一般化線形モデル」の枠組で H-NS 結合強度の回帰分析を行い、それぞれの文字列の係数パラメータを比較することで、文字列の結合強度に対する寄与の度合いを調べる。また、1 bp から 6 bp のどの長さが、結合強度を最も良く説明するのかを、情報量基準に基づいたモデル選択を行うことで判断する。以上の方法では、6 bp までの長さのあらゆる配列に関して、高親和性配列であるか否かの判断が理論的には可能であるが、一方で、ATGC 文字列として表現が不可能な「AT 含量」のような配列特徴を検出することはできない。仮に高親和性配列は A と T のどちらであっても良く、ある領域では A の連続塩基が、また別の領域では T の連続塩基が H-NS の結合強度に寄与していた場合、モデル推定の結果からは、それらの配列の両方が単一の領域に配置されていることが必要なのか、あるいはどちらであっても良いのかを区別することができない。そこで本研究では、「AT 含量」のような特徴を配列として表現するために、回帰分析に先立って、4 文字（ATGC）でコードされている DNA 配列を、2 文字でコードされた配列に変換した学習セッ

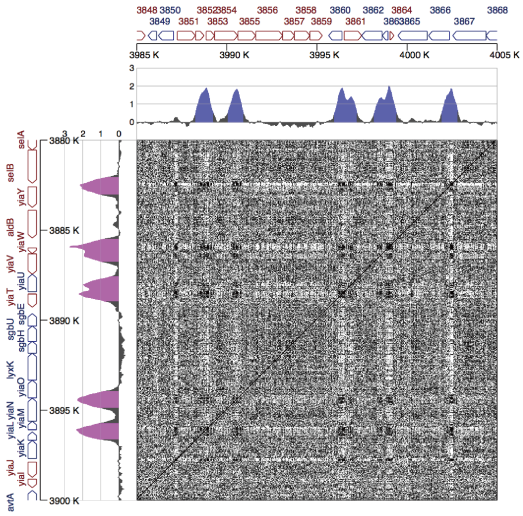

図 4.2: WS 配列によるドットマトリクス

トを用意し、それらを用いたモデルについても計算する。この変換では、配列上の A または T の文字は W（weak pairs）の 1 文字に変換され、G または C の文字は S（strong pairs）の 1 文字に変換される。この変換によって、たとえば「AT 含量」という特徴量は、WS 配列の 1 bp 変数として、モデル化することが可能となる。

このような変換の必要性は、第 2 章で導入したドットマトリクス法による H-NS 結合領域の観察によっても支持される。たとえば、図 4.2 は、事前にそれぞれの DNA 配列を WS 二文字配列に変換してから計算したドットマトリクスであるが、それぞれの株の H-NS 結合領域で、WS 二文字配列のなんらかの要素が頻出していることが見て取れる。さらに、このような特徴は、DNA4 文字配列としての保存性が失われている領域においても観察された（図 4.3）。したがって、H-NS 結合領域の配列をモデル化する際には、DNA の 4 文字配列だけでなく、WS の 2 文字配列についてもモデルを構築し、それらの間でもっとも H-NS 結合強度の説明能力が高いモデルはどれか、比較検討しなければならない。

本章の構成は次の通りである。4.2 節では、H-NS 結合領域中の結合ピーク位置を検出する手法、一般化線形モデルによる回帰分析の手法、モデル選択の手法、それぞれについて述べる。4.3 節では、これらの結果をまとめる。4.4 節では、以上の結果と、第 2 章、第 3 章で得られた結果を統合し、H-NS 結合領域が保存されながらも、H-NS 結合配列が多様化していく背景にあるメカニズムについて議論する。

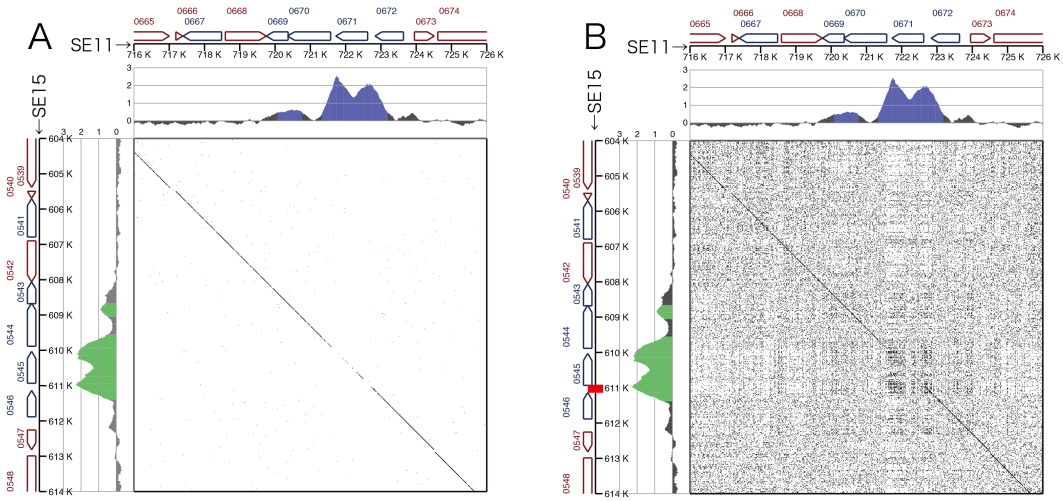

図 4.3: DNA 配列と WS 配列によるドットマトリクスの比較　(A)DNA4 文字配列で計算されたドットマトリクス。H-NS 結合領域下で配列の保存性が失われている。 (B) 同じ領域について WS2 文字配列で計算されたドットマトリクス。DNA 配列で保存性が失われていた領域内でなんらかの WS 文字要素が頻出している

4.2 手法

4.2.1 H-NS 結合ピーク位置の検出

H-NS 結合領域のいくつかは、領域中で結合強度が最大値をとる結合ピーク位置と、そこを中心としてほぼ対称に、なだらかに減衰する結合強度から構成された、長さが 1 kbp ほどの単峰型のパターンとなっている（図 2.7）。また、数 kbp の長さに広がった結合領域についても、そのような単峰型の分布がいくつか連続し、重ね合わさったもののように見える。本研究では、単峰型の結合強度パターンが、ピーク位置を中心とした正規分布に従っていると仮定する。また、多峰型の広い結合領域の結合強度については、様々な中心と分散を持ったいくつかの正規分布の重ね合わせ、すなわち混合正規分布に従った確率変数であると仮定する。

SE11 株の H-NS 結合領域は 506 個、SE15 株の H-NS 結合領域は 510 個、K-12 株の H-NS 結合領域は 436 個、それぞれのゲノム上に存在した（表 2.2）。それぞれの領域について個別に、混合正規分布モデルのパラメータフィッティングを行った。パラメータは、混合される各正規分布の平均（μ）と分散（σ）である。各正

規分布の平均パラメータが、領域内の結合強度ピーク位置を意味する。ピーク位置の検出のみを目的とするのであれば、推定するパラメータは正規分布の平均パラメータのみとし、分散パラメータは単独の H-NS 結合領域から推定した特定の値に固定する方がいいかもしれない。しかし本研究に関しては、以下の理由から分散パラメータの固定は妥当ではないと考える。H-NS 結合領域のすそ野の広がりには、テクニカルな影響とバイオロジカルな影響の 2 つがある。テクニカルな面に関しては、本研究の ChIP-seq 実験においてシーケンスのライブラリを 200 bp に調整しているために、H-NS 結合の前後の領域でマッピングされるシーケンスのカバレッジになだらかなすそ野が生じる点が考えられる。しかし、単独の H-NS 結合領域であっても、実際に観測されるすそ野の広がりには領域ごとにばらつきがある。これにはおそらくバイオロジカルな原因が影響しており、H-NS はまず親和性の高い配列に結合し、その両端で H-NS 多量体を形成しながら広がっていくと考えられているため、結合領域の「中心」と「端」には結合の安定性に差がある可能性が考えらる。つまり、安定性の高い領域が広い場合はすそ野の広がりが大きく、安定性の高い領域が狭い場合は結合が鋭いピークとして観察されることが推測される。したがって、正規分布の分散パラメータは特定の値に固定するのではなく、領域ごとに適応的に決定する必要がある。

パラメータは、混合正規分布モデルの最適化に一般的に用いられる EM アルゴリズムによって最適化を行い推定した。混合正規分布モデルの最適化計算では、混合数（いくつの正規分布を混合させるか）は、データから推定することはできず、事前に与えなければならない。しかし、それぞれの結合領域がいくつの正規分布の混合から構成されているのか（それぞれの結合領域にいくつの結合ピーク位置が含まれているのか）は未知である。そこで、各領域の混合数を 1 から 20 と設定して、それぞれの尤度を計算し、その値と、混合正規分布のパラメータ数、およびサンプル数（ここでは領域内の結合強度の和とする）から、ベイズ情報量基準（Bayesian information criterion; BIC） (Schwarz et al. 1978) を計算した。BIC は、AIC（Akaike's information criterion; 赤池情報量基準）と類似した情報量基準の一種であり、統計モデルの選択に用いられる指標である。BIC は以下で定義される。

$$BIC = -2 \ln \hat{L} + k \ln n \tag{4.1}$$

ここで、k はモデルに含まれる自由パラメータの数、n はサンプル数、$\hat{L}$ は、パラメータ θ を持つモデル M のもとで、尤度関数 $L = p(x|\theta, M)$ を最大化するパラメータ $\hat{\theta}$ を特定した際の尤度関数の値である（$\hat{L} = p(x|\hat{\theta}, M)$）。AIC と同様、BIC は、モデルの尤度と共にモデルのパラメータ数を考慮することによって、学習セットへの当てはまりの良さだけを評価するのではなく、モデルの複雑さを適切に制御し、汎化能力の高いモデルを選択する。BIC は情報理論に基づいたモデル選択基準である「最小記述長」（Minimum description length; MDL）と数学的に等価であり、AIC と比較して、パラメータ数の増加に対してのペナルティが大きいことを特徴とする。それぞれの混合数における BIC を比較し、BIC が最小となる混合数を、その領域の混合数パラメータとして採用した。最終的に、その混合数で各正規分布の平均パラメータおよび分散パラメータの最適化を行い、平均パラメータの位置をその領域のピーク位置として決定した。最適化のための EM アルゴリズム、BIC の計算には、Python 言語で実装された機械学習の統合パッケージである scikit-learn (Pedregosa et al. 2011) の sklearn.mixture.GMM を用いた。

4.2.2 ガンマ一般化線形モデルによるモデル化

前項の操作で取得したピーク位置のそれぞれについて、前後 100 bp、計 200 bp の長さの配列と、その位置における H-NS 結合強度の情報を取得した。本項の解析の目的は、ある DNA 配列が与えられたときに、その配列上の平均 H-NS 結合強度を予測する統計モデルを構築することである。本項で用いる記号を表 4.1 にまとめる。

表 4.1: 本項で用いる記号

記号	説明
y	H-NS 結合強度
i	データセット（ピーク位置周辺配列）の添字
n	データセットの総数
k	特定の k-mer 文字列の添字
x_k	特定の k-mer 文字列
$\mathbf{x}$	k-mer 文字列集合
$f(x_k)$	特定の k-mer 文字列の出現回数/出現期待値
w_k	特定の k-mer 文字列の重み係数
σ^2	正規分布の分散パラメータ
a	ガンマ分布の形状パラメータ
b	ガンマ分布の尺度パラメータ
$\boldsymbol{W_a}$	ガンマ分布の a パラメータを決定する各 k-mer の重みベクトル
$\boldsymbol{W_b}$	ガンマ分布の b パラメータを決定する各 k-mer の重みベクトル

上述の全ピーク位置周辺配列について、学習セットを次のように構築した。

1. 応答変数：200 bp 内平均 H-NS 結合強度
2. 説明変数:200 bp の DNA 配列に含まれる k-mer（k 連続塩基、$k \in \{1, 2, 3, 4, 5, 6\}$）の出現回数／出現期待値

たとえば、ATGC の 4 文字配列における 1-mer 予測モデルの場合、説明変数は、配列に含まれる A, T, G, C それぞれの文字をカウントし、期待値（それぞれが等確率で出現すると仮定する。200 bp 配列中の出現期待値はそれぞれ 50 回ずつ）で割った値となる。

また、緒言で述べたように DNA の ATGC4 文字配列は、H-NS 結合領域の説明変数としては適切ではない可能性がある（「AT 含量」といった特徴を表現できない）ため、ここではさらに、200bp の WS 配列（A および T を W の 1 文字に、G および C を S の 1 文字に変換した配列）に関する k-mer（$k \in \{1, 2, 3, 4, 5, 6\}$）を説明変数としたモデルについても比較する。

一般的な線形重回帰分析の場合、説明変数と応答変数の変換には恒等リンク関数を設定し、応答変数の従う分布は正規分布としたモデルを考える。つまり、$\boldsymbol{x}$ を

k-mer の文字列集合、y を応答変数（平均 H-NS 結合強度）とし、$f(\boldsymbol{x})$ を各 k-mer 文字列の出現回数／出現期待値、$\boldsymbol{w}$ を各 k-mer 文字列の係数パラメータ、C を切片の定数パラメータとすると、

$$y = \sum_k w_k f(x_k) + C + \epsilon \tag{4.2}$$

$$\epsilon \sim N(0, \sigma^2) \tag{4.3}$$

となる。ただし N は平均 0、分散 σ^2 の正規分布とする。定数パラメータを説明変数の中にまとめてしまうと、結局これは、応答変数 y が、平均を $\sum_i w_i f(x_i)$、分散を σ^2 とする正規分布に従うことになるので、y の確率密度関数を $p(y|\boldsymbol{x}, \boldsymbol{w}, \sigma)$ と書くと、

$$p(y|\boldsymbol{x}, \boldsymbol{w}, \sigma) = N(\sum_k w_k f(x_k), \sigma^2) \tag{4.4}$$

であり、学習セット 1 〜 n について対数尤度は、

$$\sum_{i=1}^{n} log(p(y_i|\boldsymbol{x}, \boldsymbol{w}, \sigma)) \tag{4.5}$$

となる。したがって、求める係数パラメータ $\hat{\boldsymbol{w}}$ は、

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} \left(\sum_{i=1}^{n} log(p(y_i|\boldsymbol{x}, \boldsymbol{w}, \sigma)) \right) \tag{4.6}$$

となる。

リンク関数に恒等写像、応答変数の確率分布に正規分布を仮定する線形重回帰モデルは、非常に制約の強いモデルであり、次の 2 点が要求される。

1. 応答変数と説明変数の関係性が線形である
2. 応答変数が正規分布に従って分布する、すなわち応答変数が負の値をとりうる

だが、本研究で扱うデータについて、これらの制約は不自然である。まず 1 について、説明変数と応答変数間の線形性は自明ではない。むしろ、説明変数の取り得る値の範囲が限定されている（200 bp 内のカウントに依存する）一方で、応答変数の取り得る値の範囲が大きい（数十から数千）ことを考えると、それらの間

に線形性は仮定できないと考える方が自然である。また 2 に関しては、本研究で扱うデータの統計モデルとして不適切である。本研究では、応答変数は平均 H-NS 結合強度であるため非負値であるが、正規分布は $-\infty$ から ∞ の範囲の確率変数上に定義される確率分布である。つまり、以上のモデルは、負の値を予測分布に含むことになり、予測モデルとして適当なものとはなり得ない。さらに、上記の線形重回帰モデルでは、結合強度値のすべての範囲において誤差が等分散正規分布に従うことを仮定しているが、これも自明ではなく、おそらく適切な仮定ではない。

そこで本研究では、応答変数の誤差の分布がガンマ分布にしたがうと仮定した、ガンマ一般化線形モデルを構築する。ガンマ分布は連続値確率分布の一種であり、台は正の実数である。つまり確率変数はゼロ未満の値をとらない。ガンマ分布の確率密度関数は 2 つのパラメータ a と b を用いて、

$$p(y) = y^{a-1}\frac{e^{-y/b}}{\Gamma(a)b^a} \tag{4.7}$$

と定義される。ここで $\Gamma()$ はガンマ関数である。a、b はそれぞれガンマ分布の形状パラメータと尺度パラメータであり、正の値をとる。ガンマ分布の平均と分散はパラメータ a,b を用いてそれぞれ、$E(y) = ab, V(y) = ab^2$ となる。ガンマ分布の確率密度関数のパラメータ a、b に対する依存性を図 4.4 に示す。パラメータ a,b どちらも分布の平均値に影響を及ぼすが、パラメータ b は分布の分散に及ぼす影響が大きいことが特徴である。

ここで、それぞれの観測値（H-NS 結合強度）の k-mer への依存性を表現するために、観測値 y_i がしたがうガンマ分布のパラメータ $a(i), b(i)$ を次のように定義する。

$$a(i) = \exp(\boldsymbol{W_a}^T f(\boldsymbol{x_i})) \tag{4.8}$$

$$b(i) = \exp(\boldsymbol{W_b}^T f(\boldsymbol{x_i})) \tag{4.9}$$

ただし、$\boldsymbol{f}(x_i)$ は観測値 y_i に対応する配列の k-mer 特徴量、$\boldsymbol{W_a}$ はパラメータ a に対する k-mer 特徴量の重みベクトル、$\boldsymbol{W_b}$ はパラメータ b に対する k-mer 特徴量の重みベクトルである。すなわち、一般化線形モデルのリンク関数として対数リンク関数を採用する。ガンマ分布のパラメータ a, b は正の実数であるが、特徴量を指数の肩に乗せることでこの制約は満たされている。

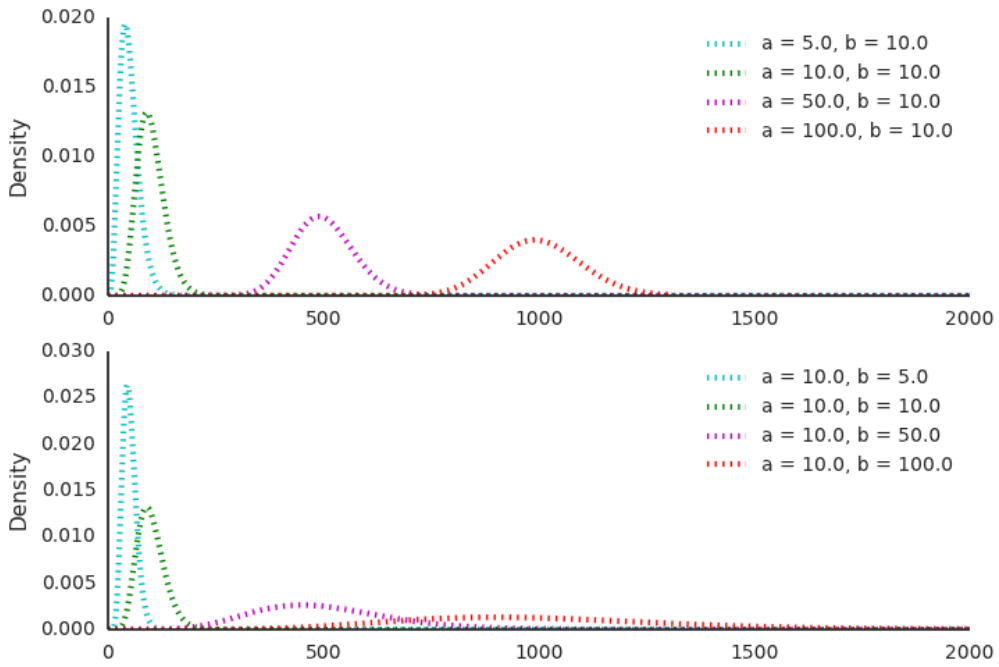

図 4.4: ガンマ分布は a、b の 2 つのパラメータで特徴付けられる。上はパラメータ b を固定し a を変化させたときの確率密度関数の形状の変化、下はパラメータ a を固定し b を変化させたときの形状の変化である。

観測値全体にわたる尤度関数は、

$$\prod_{i=1}^{n} p(y_i|\boldsymbol{W_a}, \boldsymbol{W_b}) \tag{4.10}$$

となり、対数尤度関数は、

$$\begin{aligned}&\sum_{i=1}^{n} log(p(y_i|\boldsymbol{W_a}, \boldsymbol{W_b}))\\&\quad= \sum_{i=1}^{n}((a(i)-1)log(y_i) - log(\Gamma(a(i))) - a(i)log(b(i)) - \frac{y_i}{b})\end{aligned} \tag{4.11}$$

となる。この対数尤度関数が最大となるように、特徴量の重みベクトル W_a および W_b を決定する。本研究におけるガンマ一般化線形モデルではパラメータが高次元（数十から数百次元）であるため、勾配法に基づく数値最適化では局所最適解に陥

りやすい。したがって、本研究ではパラメータ空間を、Metropolis-Hastings アルゴリズム（M-H 法）に基づくマルコフ連鎖モンテカルロ法（Markov chain Monte Carlo methods; MCMC）によって確率的に探索するプログラムを実装した。各パラメータについて正規分布に基づくランダムウォークを繰り返し、各ステップで上記の尤度関数を評価し、その比を M-H 法の採択確率として扱う。なお、パラメータについて L2 正則化を導入する。これは、本節のモデルをベイズモデルと見た場合、それぞれのパラメータの事前分布がゼロを中心とする正規分布であると仮定していることに対応する。

ATGC4 文字配列の k-mer モデル（$k \in \{1, 2, 3, 4, 5, 6\}$）および WS2 文字配列の k-mer モデル（$k \in \{1, 2, 3, 4, 5, 6\}$）についてそれぞれ、ガンマ一般化線形モデルのパラメータ最適化計算を MCMC によって行った。MCMC の収束を確認した上で、最終的に得られたパラメータセットに基づく尤度関数を計算し、モデルのベイズ情報量基準を計算した。もっとも汎化能力の高いモデルは、ベイズ情報量基準が最小となるモデルとして決定した。最後に、得られたモデルのパラメータを比較することによって、H-NS 結合領域の安定性にもっとも寄与している配列要素を明らかにした。

4.3 結果

4.3.1 混合正規分布モデルによる H-NS 結合領域のモデル化およびピーク位置検出

3 株の H-NS 結合領域それぞれについて、混合正規分布モデルによってモデル化し、正規分布の平均パラメータの座標として H-NS 結合ピーク位置を検出した。正規分布の混合数に関して各領域で 1 から 20 個のモデルの最適化計算を行い、BIC が最小になるモデルを決定した。結果、SE11 株については 2,231、SE15 株については 2,286、K-12 株については 2,151 の H-NS 結合ピーク位置を検出した。H-NS 結合領域の混合正規分布によるモデル化の例を図 4.5 に示す。

いくつか、H-NS 結合強度の波形が正規分布から逸脱した形状の場合に、正規分布によるモデル化が適切でないように思われる領域があるが、多くの領域でおおむね、直感的なピーク位置と正規分布の平均パラメータの座標が一致している。

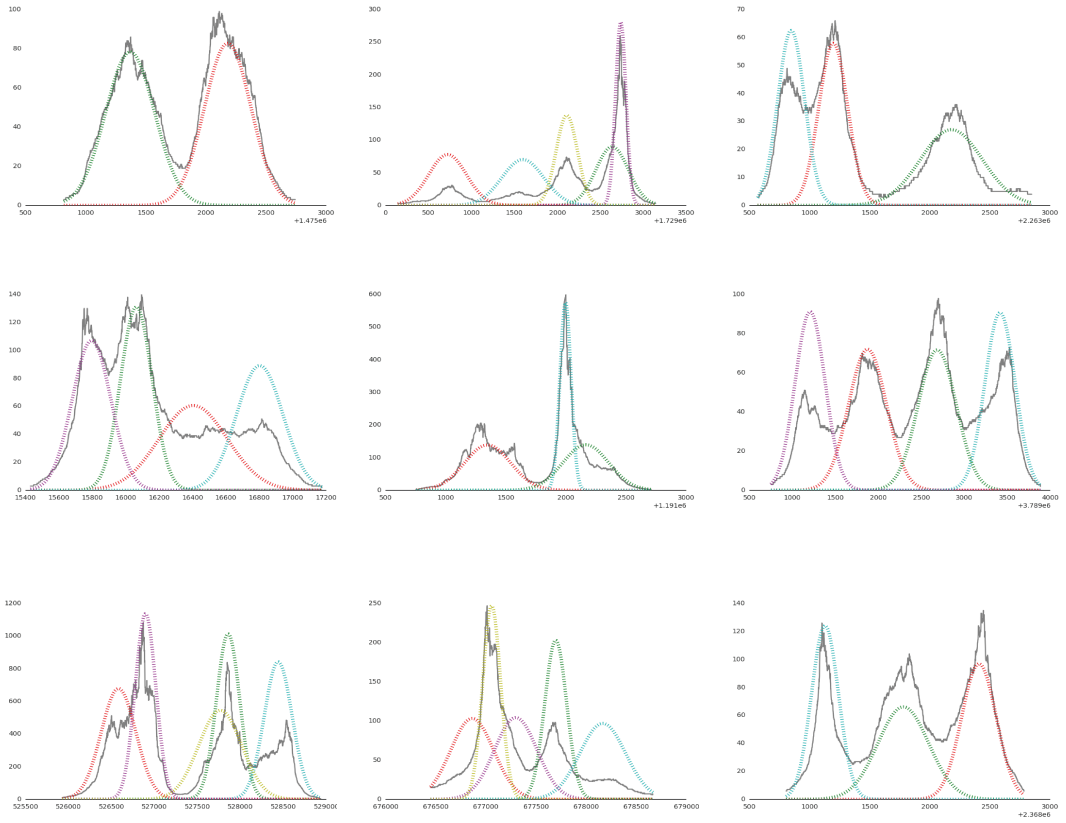

図 4.5: 混合正規分布による H-NS 結合領域のモデル化　上側 3 枚が SE11 株、中央 3 枚が SE15 株、下側 3 枚が K-12 株における推定結果の例。いずれも、灰色の線が ChAP-seq 実験によって得られた H-NS 結合強度波形、異なる色で描かれている線がそれぞれ推定された正規分布を示している

検出されたピーク位置について、前後 100bp の配列を切り出し、ガンマ一般化線形モデルの学習セットを構築するが、もしピーク位置の H-NS 結合強度に染色体上の位置依存的な傾向が見られたならば、H-NS 結合強度が本質的に DNA 配列に依存すると仮定している本章のモデル化が妥当ではない可能性がある。したがって、ガンマ一般化線形モデルによる学習に先立って、H-NS 結合ピーク位置の染色体上の位置と、その位置における H-NS 結合強度の関係を調べた。まず、それぞれの株の染色体の Ori ドメイン（複製開始領域）を DoriC データベース (Gao and Zhang 2007, Gao et al. 2013) より取得した。それぞれの株の Ter ドメイン（複製終結領域）は、Ori ドメインから染色体上の位置でちょうど反対に位置する領域（角度座標では π ラジアン）として推定した。その後、それぞれの H-NS 結合ピーク位置を Ori からの Ter への角度座標に変換して、Ori からの角度と H-NS 結合強度の関

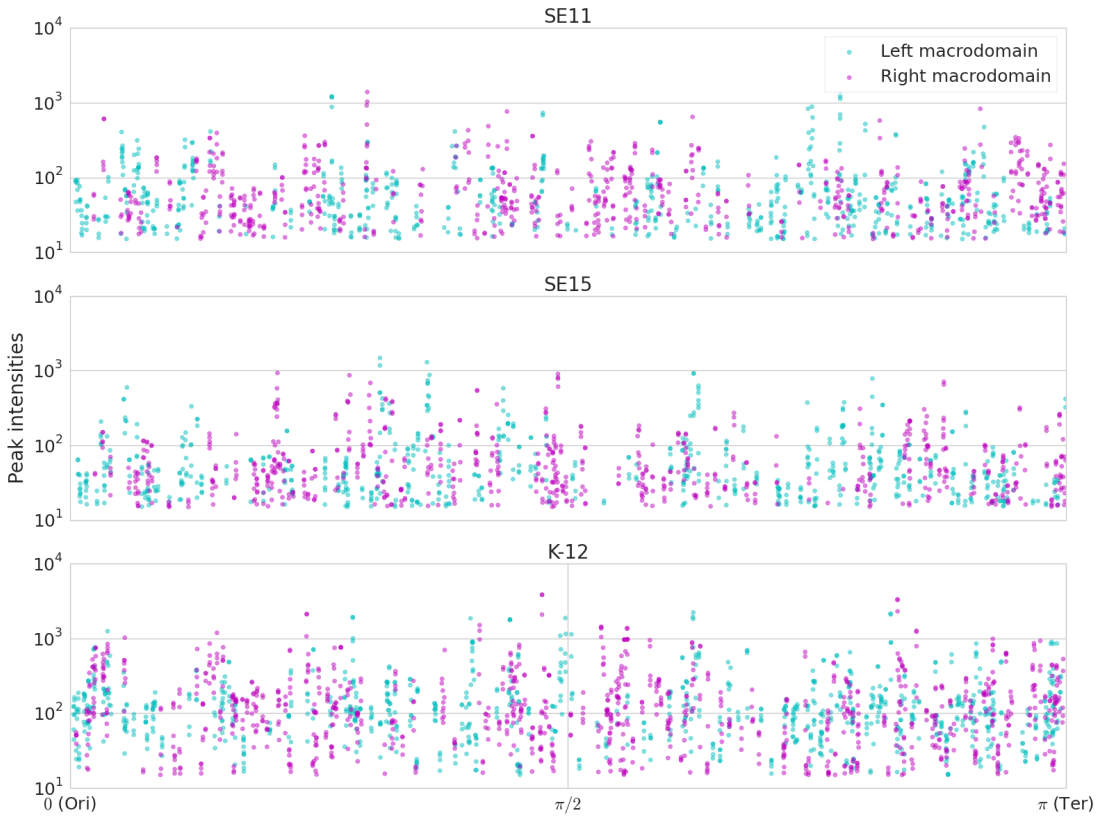

図 4.6: 染色体上の H-NS 結合強度の分布　各プロットが推定された H-NS 結合ピークを示している。x 軸が染色体上の位置（左端が Ori、右端が Ter としたときの染色体上の角度座標）、y 軸が対応するピークの H-NS 結合強度を示している。さらに、水色が染色体の Left macrodomain, 紫色が染色体の Right macrodomain 上のピークである。上が SE11 株、中央が SE15 株、下が K-12 株における H-NS 結合ピーク位置の分布を示している

係を調べた。結果を図 4.6 に示す。

H-NS 結合ピーク位置は 3 株のいずれにおいても、染色体上の Ori ドメイン（複製開始領域）から Ter ドメイン（複製終結領域）まで広く分布し、それらの間で特に H-NS 結合強度の偏りは見られなかった。また、染色体上の左マクロドメイン、右マクロドメイン間での H-NS 結合強度の偏りも見られなかった。以上の結果から、染色体上の領域依存的な H-NS 結合強度の偏りは見られず、H-NS 結合強度は結合領域下の DNA 配列に依存して決定されると仮定した本章のモデルは妥当であると考えられる。

4.3.2 ガンマ一般化線形モデルによる H-NS 結合強度のモデル化

抽出された H-NS 結合ピーク位置情報を用いて、それぞれのピーク位置の前後 100bp の配列を切り出し、それぞれの配列について、ATGC4 文字配列の k-mer（$k \in \{1,2,3,4,5,6\}$）および、WS2 文字配列に変換した場合の k-mer（$k \in \{1,2,3,4,5,6\}$）の存在量を特徴量とする特徴ベクトルを構成して、それぞれの領域の平均 H-NS 結合強度を応答変数とするガンマ一般化線形モデルの学習を行った。ATGC4 文字配列について 1-mer から 6-mer の 6 つのモデル、WS2 文字配列について 1-mer から 6-mer の 6 つのモデル、計 12 個のモデルのパラメータをそれぞれ MCMC で最適化して決定し、得られた 12 個のモデルをベイズ情報量基準によって比較した。結果を図 4.7 に示す。

ベイズ情報量基準がもっとも小さいモデルが、モデルの複雑さとデータへの適合度がもっともバランスのとれた、汎化能力の高いモデルである。3 株すべてで同じ結果が得られ、ベイズ情報量基準が最小となったのは、WS2 文字配列の 5-mer (5 連続塩基）を説明変数とするモデルであった（図 4.7)。これらの 3 株はそれぞれ独立な学習データセットを構築して独立にモデルの学習を行ったので、3 株でベイズ情報量基準が最小となるモデルが一致していることは、これらの学習結果がデータセットの違いに対してロバストであることを意味している。

WS 5-mer のモデルで学習されたパラメータによって予測された H-NS 結合強度と、実際の観測値との比較を図 4.8 に示す。

決定係数はそれぞれ、0.24（SE11 株）、0.25（SE15 株）、0.26（K-12 株）であった。プロットをよく見ると、特に H-NS 結合強度が本来高い領域に関して予測精度が低く、結合強度を過小に予測している。H-NS 結合強度の高い領域は、実験に用いた個体の多くで結合が見られる、安定した H-NS 結合領域である。このような領域では、多量体形成による安定化効果、あるいは H-NS のパラログである StpA (Uyar et al. 2009) や、一部の結合領域で H-NS/StpA とヘテロダイマーを形成し、H-NS の結合の安定性に寄与していると考えられている Hha および YdgT など (Ueda et al. 2013) のタンパク質との協調効果によって、H-NS 結合強度が配列によって規定される以上に高くなっているのかもしれない。

最後に、ベイズ情報量基準最小となった WS2 文字配列 5-mer モデルについて、学習された各 5-mer の重みパラメータの値を調べた。ガンマ分布はパラメータ a および b の値がそれぞれ大きいほど確率変数が大きな値を中心として分布するた

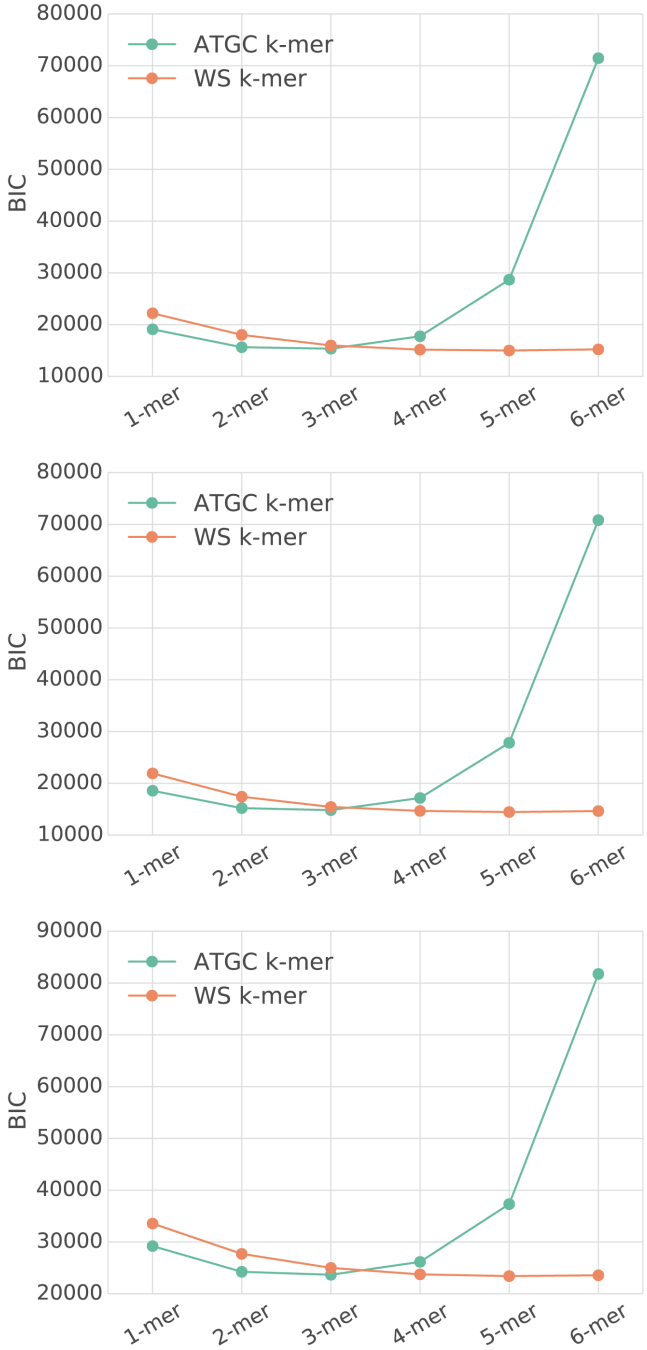

図 4.7: ベイズ情報量基準によるモデル比較　上から SE11 株、SE15 株、K-12 株の結果。緑色のプロットは ATGC4 文字配列によるモデルの BIC を、オレンジ色のプロットは WS2 文字配列によるモデルの BIC を示している

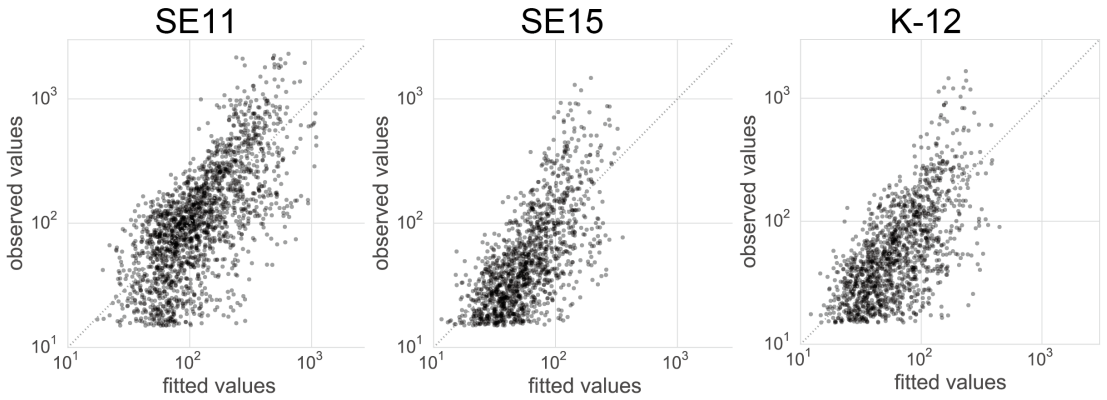

図 4.8: モデルの予測値と観測値の比較　左から SE11 株、SE15 株、K-12 株の結果を示す。x 軸はモデルによって予測された H-NS 結合強度、y 軸は実際の H-NS 結合強度。

め、パラメータ a および b に指数的に作用する重みベクトル W_a, W_b の値が大きいほど、その k-mer の存在量が H-NS 結合強度を大きくする方向に作用しているとみなすことができる。結果を図 4.9 に示す。なお、バイアスパラメータの重みは省略している。

いずれの株においても、学習された重み係数の値はほぼ同じパターンとなった。まず、パラメータ a（形状パラメータ）に関わる重み係数 W_a は、各 WS 5-mer で大きな差はなかった。パラメータ b（尺度パラメータ）に関わる重み係数 W_b の値は各 WS 5-mer で大きな差があり、いずれの株においても、WWWWW（W=A または T の 5 連続配列要素）が最も大きな係数となっていた（図 4.9 右端）。したがって、H-NS 結合領域内においては、WWWWW という 5 連続の配列要素が多く存在するほど H-NS 結合強度が大きくなる、という学習結果が得られた。係数の値としては WWWWW がきわだって大きいが、次に大きな係数の値をとっていたのは、5-mer のうちの 4 つが W で占められる配列要素（SWWWW, WSWWW, WWSWW, WWWSW, WWWWS）であった。

k-mer 特徴量にかかる重み係数は、特定の k-mer の H-NS 結合強度に対する相対的な「貢献度合い」と解釈することができる。重みの大きな k-mer は、その k-mer の存在量がガンマ分布パラメータの増加に大きく寄与するが、重みの小さな k-mer は、配列上の存在量の多寡がガンマ分布パラメータにほとんど影響を与えない。ガンマ分布の平均と分散は 2 つのパラメータ a, b によって表され、それぞれ

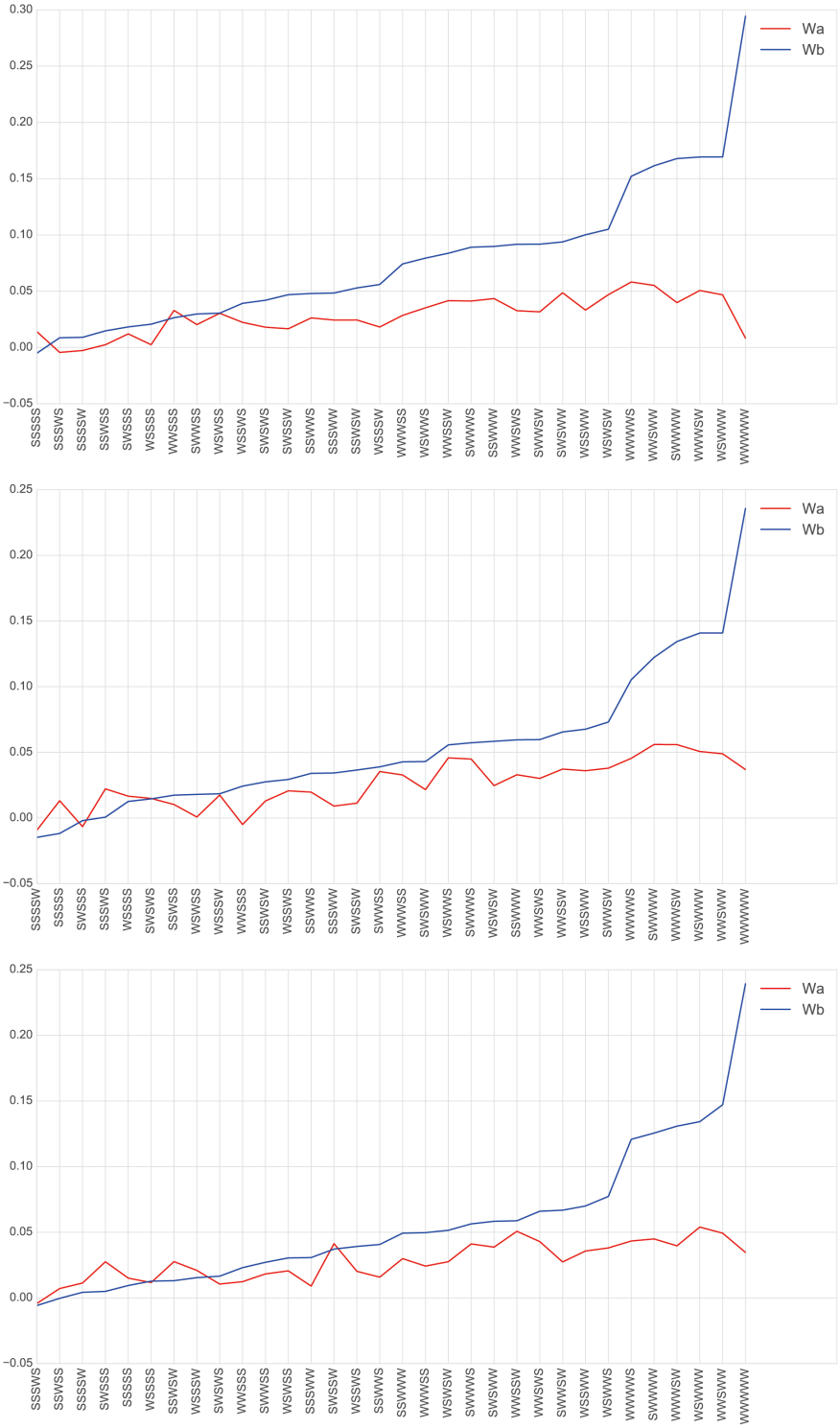

図 4.9: WS 5-mer モデルで学習された重みベクトルの係数　上から SE11 株、SE15 株、K-12 株の結果を示す。x 軸は各 WS 5-mer 配列要素、y 軸はそれぞれの重み係数の値、赤線は W_a の各値を、青線は W_b の各値を示している

$E(y) = ab, V(y) = ab^2$ となる。したがって、パラメータ b はガンマ分布からサンプルされる値のばらつきへの影響がより大きい。パラメータ a に関わる重み係数 W_a に関して k-mer 間で差が見られなかった一方、パラメータ b に関わる重み係数 W_b で k-mer ごとの差が見られたことから、パラメータ b が H-NS 結合領域ごとに大きく変化する学習結果が得られた。この結果は、「配列上の k-mer 特徴の影響で H-NS 結合強度が高く観測され得る領域では、実際の観測値は大きくばらつく傾向があり、H-NS 結合強度が低いはずの領域では、安定して低い」モデルが学習されたことを意味している。H-NS 結合強度分布のこのようなパターンは実際の観測値においても見られ（図 2.3B）、再現性実験間の両対数グラフにおいて、直線的なプロットの広がりがすべてのスケールでほぼ同一であることは、H-NS 結合強度が高い領域で大きくばらつきやすいことを示している。したがって、パラメータ b に影響する重みに関して k-mer ごとの差が大きくなったことは、H-NS 結合強度のこのような特徴を学習した結果であると考えられる。

以上の結果から、ゲノム上の特定の領域において、W（A または T）が局所的に連続した配列要素がある程度の長さの範囲内で密集していることが、H-NS の結合を安定化させるために必要であり、そのような配列パターンは、その領域の H-NS 結合強度を予測するのに十分な情報を持っていることがわかった。

一方で、W 連続配列を構成する ATGC4 文字配列が、特定の配列（たとえば ATA など）に限定されている可能性はある。ベイズ情報量基準によるモデル比較の結果（図 4.7）、ATGC4 文字配列の 3-mer モデルについて、WS 5-mer モデルと同程度に小さなベイズ情報量基準となっていた。そこで、ATGC 3-mer のモデルで学習された重み係数について調べてみると（図 4.10）、K-12 株においては“ ATA ”が大きな係数として得られているが、結果は株間で安定せず、SE11 株においては“ TTT ”、“ AAA ”、“ ATA ”、“ TAT ”、“ TAA ”が同程度に、SE15 株においては“ ATA ”と“ TTT ”が同程度に大きな係数となっていた。きわだって寄与の大きい特定の ATGC 配列要素は存在せず、同程度の大きさの係数でいくつかの配列要素にばらけていることから、それら単独での H-NS 結合強度の予測能力はそれほど大きくなく、やはり A または T を区別しない W の連続配列要素の存在量によって H-NS 結合強度が規定されている可能性が高いと考えられる。

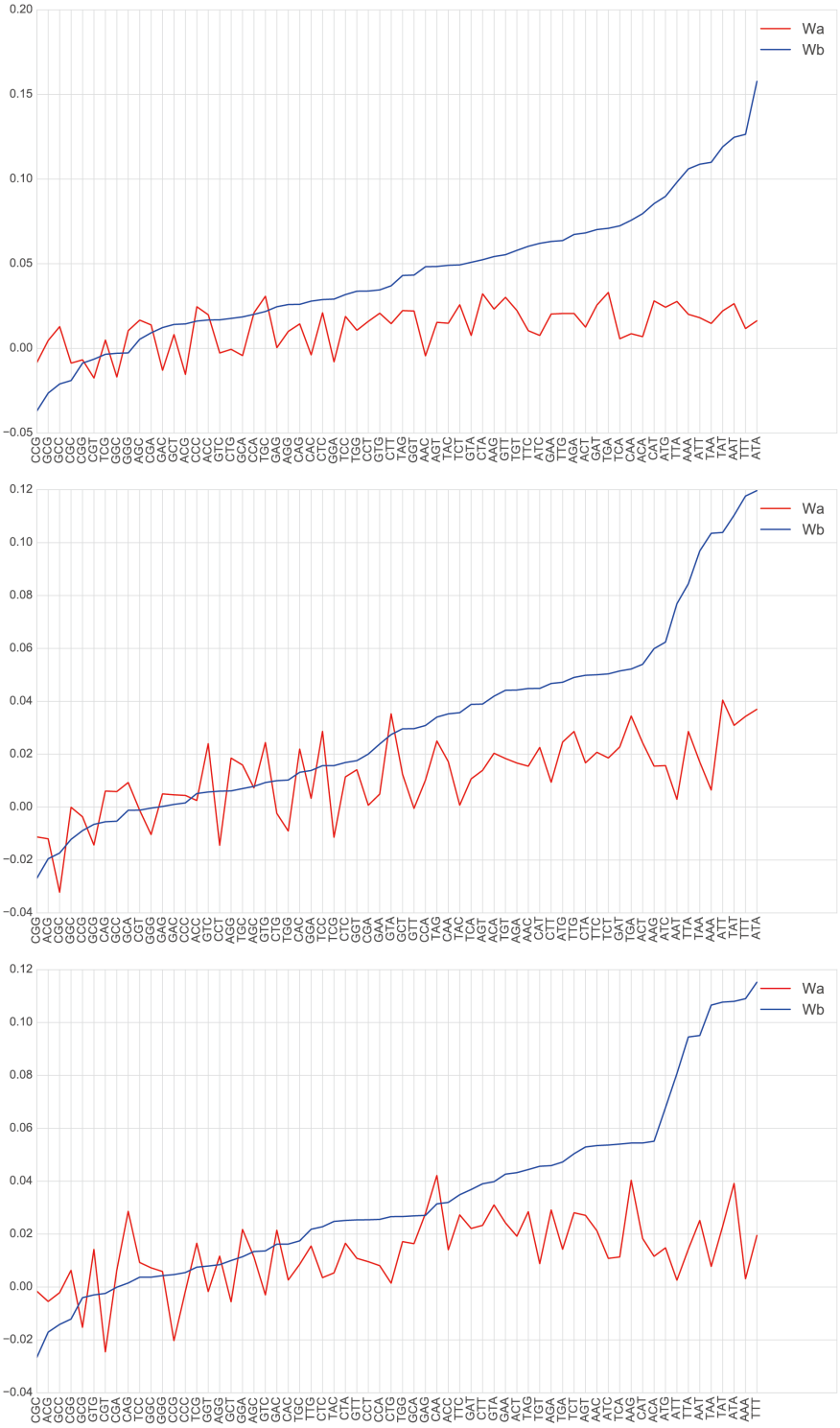

図 4.10: ATGC 3-mer モデルで学習された重みベクトルの係数　上から SE11 株、SE15 株、K-12 株の結果を示す。x 軸は各 ATGC 3-mer 配列要素、y 軸はそれぞれの重み係数の値、赤線は W_a の各値を、青線は W_b の各値を示している

4.4 考察

本章では、H-NS 結合ピーク位置の抽出と、その周辺配列における配列要素（k-mer）を説明変数とした H-NS 結合強度のガンマ一般化線形モデル構築によって、以下の 2 点を明らかにした。

1. H-NS 結合強度は、A,T,G,C の 4 文字配列ではなく、A および T を W に、G および C を S に変換した WS2 文字配列の 5-mer 配列要素によってもっともよく説明される

2. WS 5-mer 配列要素の中では、W5 連続配列（WWWWW）の結合強度への寄与がもっとも大きい

この結果は、これまで指摘されてきた「AT 含量の高さ」によって H-NS の結合を説明するモデルと矛盾するものではないが、意味するところはわずかに異なる。実際、「AT 含量の高さ」で H-NS 結合強度が規定されるモデルは、本章で扱ったモデルの中では WS 1-mer モデルがそれに該当するが、その結合強度予測能力は高くなかった（図 4.7）。WS 5-mer モデルの重み係数比較の結果から、H-NS の結合にもっとも寄与する配列要素は W5 連続配列であることがわかり、したがって結果的に、H-NS 結合強度が高い領域において配列中の AT 含量（W 含量）が高く観測されるが、おそらく AT 含量の高さ以上に、「W が連続している配列」が密集していることが H-NS 結合の安定化において重要なのだと考えられる。

一方で、そのような特徴が保持されてさえいれば、その領域上で変異が蓄積して、ATGC4 文字配列としてのパターンが失われたとしても、H-NS の結合が維持できるのかもしれない。実際に、DNA 配列の相同性が失われても、WS2 文字配列の密集という特徴が保たれている領域が存在し（図 4.3）、また、A,T,G,C4 文字配列で学習された 3-mer のガンマ一般化線形モデルでは、きわだって大きな重み係数の値を持つ配列要素は存在しなかった。H-NS 結合領域はこのような特徴を持つことによって、第 3 章で見たように多様性が非常に高い領域であったとしても、第 2 章で見たように結合領域を長期的に維持し続けることが可能なのかもしれない。

先行研究によって、H-NS の結合を導く高親和性配列として、AT 含量 (Lucchini et al. 2006, Navarre et al. 2006)、A-tract 配列 (Crothers et al. 1990, Zuber et al. 1994, Oshima et al. 2006)、AT 含量の高い配列における TpA2 連続塩基の存在 (Kahramanoglou et al. 2011, Gordon et al. 2011) が提案されている。これら

は本章で示した結果と矛盾するものではない。多くの先行研究では、H-NS 結合配列と H-NS 非結合配列との比較によって H-NS の結合を安定化する配列の特定を試みているが、H-NS 結合配列が本来 AT 含量が高い傾向にあることから、そのような比較によっては、単に AT 含量の高い領域で確率的に出現しやすい配列を特定してしまうことになりかねない。本研究では、H-NS 結合領域のみを対象としてモデルを構築し、さらに ChAP-seq 解析によって得られた H-NS 結合強度情報を積極的に利用することによって、H-NS の結合を安定化している配列要素を特定した。ChAP-seq 解析によって得られる結合強度情報は、細菌集団中の DNA タンパク質の結合の安定性を反映した定量的な情報である。さらに H-NS は、スーパーコイルの変化や (Tobe et al. 1995)、H-NS の部分的なパラログであるタンパク質とのヘテロダイマー形成による H-NS 多量体形成の不安定化 (Williamson and Free 2005)、温度や浸透圧による抑制解除 (Prosseda et al. 2004) など、他の様々な要素と相互作用しながら転写制御を行い (Stoebel et al. 2008)、またそれらの要素は結合領域によってそれぞれ異なっている。つまり、H-NS は「結合しているか否か」というデジタルなシグナルを用いて転写制御を行っているのではなく、H-NS がいかに安定的に結合しているかというアナログな結合強度情報を用いて、つまり連続的に変化する量を連続量として解釈することによって複雑な転写制御メカニズムを実現している。したがって、H-NS の結合が特定の領域でいかに安定化しているか、すなわち H-NS 結合強度が特定の領域でどれほどの値を示しているかという情報は、*in vivo* における H-NS の安定化のメカニズムを知る上で重要だと考えられる。本章の手法はそのような結合強度情報を、結合領域下の DNA 配列情報に回帰させる目的で利用した。回帰モデルでは、いくつかの領域、特に H-NS 結合強度が非常に高い領域において予測精度が低く、結合強度が過小に推定されていた。予測から逸脱した領域は興味深い研究対象であり、そこでは DNA 配列によって規定され得ない、*in vivo* で結合安定性を高めているなんらかの要因が見いだされるかもしれない。

WS ドットマトリクスによる観察（図 4.2、図 4.3）およびガンマ一般化線形モデルによる学習結果から、H-NS 結合領域では、その他の領域と明瞭に区別される配列パターンが存在し、短い AT-rich 配列がある程度の広い範囲で高度に密集しているという環境の維持が、H-NS の結合安定性を高めるために不可欠であることが示唆された。この結果はまた、H-NS が広範囲に結合する際に、まず高親和性配

列に結合して核形成を行い、その後、H-NS 多量体を形成しながら周辺に広がっていくという、H-NS 結合の核形成・伝播モデル (Lang et al. 2007) とも一致している。短い AT-rich 配列の密集は、H-NS が DNA と周期的に相互作用し、それらの領域で核形成を行って、その後より大きな H-NS-DNA 複合体を形成する際の足場を提供するために利用されていると考えられる (Arold et al. 2010)。短い AT-rich 配列の密集という H-NS の結合条件の緩さから、H-NS と DNA との間の相互作用は非常に柔軟であるといえる。このような柔軟性によって、H-NS 結合領域上で親和性配列の獲得・欠失が生じながらも、H-NS の結合を維持し続けることが可能となり、また、H-NS がそのような領域上の遺伝子の転写を抑えることで、変異の蓄積によって生じうるネガティブな効果を回避しつつ、遺伝子の機能・制御の多様性を高めることができるのだと考えられる。したがって、H-NS 結合領域上に観察されるこのような配列的特徴もまた、Dorman の提唱した仮説 (Dorman 2007) に基づいた役割を H-NS が担うために必要な要素であるのかもしれない。

第5章　総括

本章ではここまでの結果をまとめ、本研究の結論を述べる。それを踏まえ、細菌の進化における H-NS の役割を考察する。また、いくつかの未解決問題に触れ、今後の展望を述べる。

5.1　本論文のまとめ

本研究では、進化系統的に近い大腸菌 K-12 株と SE11 株、およびそれらとは進化系統的に離れた大腸菌 SE15 株のゲノム上に結合する H-NS タンパク質の比較 ChAP-seq 解析を行い、その結果を比較ゲノム解析、分子進化解析、そして多変量解析と結びつけることによって、大腸菌の進化における H-NS の役割を議論してきた。

遺伝子水平伝播（HGT）は、細菌、特に大腸菌が様々な環境に適応する上で非常に重要な進化のメカニズムであるが、外来性遺伝子は既存のシステムに擾乱を引き起こし、生存にとって大きなコストとなりうる。したがって HGT によって多様性を獲得している細菌は、外来性遺伝子の獲得初期段階においてそれらの発現を抑制し、安定的にゲノムへと取り込む何らかのシステムを有しているはずである。近年、H-NS と呼ばれるタンパク質がその役割の一端を担っていることが明らかとなった。H-NS は大腸菌ゲノムの高次構造形成に関わる核様体タンパク質の一種として特定された DNA 結合タンパク質であるが、外来性遺伝子を特異的に認識し、その転写を抑制することで外来性遺伝子の発現による有害な効果を低減する機能を持つことが示された。しかし、HGT はけっして外来性遺伝子の宿主ゲノムへの組み込みによって完結する現象ではなく、HGT の完遂には異物である外来性遺伝子を宿主の転写調節ネットワークに適応させていくために、進化のスケールで非常に長い時間が必要となる。外来性遺伝子はいかにして宿主のシステムに組み込まれるのか？外来性遺伝子の適応プロセスに、H-NS はどのように関わってきたのか？大腸菌の進化の観点で、H-NS が外来性遺伝子の進化、多様化、および宿

主ゲノムへの適応にどのような役割を果たしてきたのか？これらの問題はこれまでほとんど明らかにされていない。

本研究では、大腸菌株間の H-NS の結合、配列の多様性、および配列パターンを徹底的に比較することによって、上記の問題の解決に取り組んだ。ここで、序論で挙げた 3 つの問題点に対応させるかたちで、各章の結果をまとめる。

5.1.1 結果のまとめ

H-NS の結合は進化の過程で保存されるのか？

第 2 章では、大腸菌 3 株のゲノム上 H-NS 結合領域を網羅的に取得する ChAP-seq 解析と、3 株の比較ゲノム解析の結果を組み合わせることによって、3 株で共有されている“Common”ゲノム領域において H-NS の結合領域が高度に維持されていることを明らかにした。“Shared”領域や“Specific”領域など、比較的近年の HGT で獲得されたと考えられる外来性のゲノム領域において、先行研究で報告されているように H-NS の結合が頻繁に見られたが、“Common”領域においても H-NS の結合は多数見られ、それらの結合パターンはきわめて精確に保存されていた。また、本研究で用いた大腸菌 K-12 株および SE11 株と大腸菌 SE15 株は大腸菌系統群の中で進化的に遠く、それらの共有ゲノム領域はおそらく大腸菌系統群の共通祖先で保持されていたゲノム領域と考えられるため、少なくとも大腸菌系統群の多様化以降、H-NS の結合がそれぞれの株で維持されてきたことを示唆した。

H-NS 結合領域下の DNA 配列はどのような進化的特徴を持っているのか？

第 3 章では、大腸菌 44 株の分子進化解析によって大腸菌系統群のオーソログ遺伝子、および相同遺伝子間領域を特定し、それらを H-NS 結合・非結合に分類することによって、H-NS 結合領域に見られる進化的なパターンを見いだした。大腸菌の様々な株が分岐する以前、共通祖先が有していたと思われる遺伝子および転写調節領域において、H-NS が結合している場合、多様性が大きい傾向にあった。そのような多様性の増加は、オーソログ遺伝子の非同義座位と、転写制御領域を含む相同遺伝子間領域に限定して観察された。すなわち、純化選択の効果によって多様性が低く抑えられているはずの領域において、H-NS が結合している場合に多様性が増大する傾向がみられた。この結果から、H-NS 結合領域下では、H-NS

の結合によって純化選択の効果が減少することで、多様性が生み出されている可能性を示唆した。

H-NS 結合領域下の DNA 配列はどのような配列的特徴を持っているのか？

第 4 章では、H-NS 結合ピーク位置の抽出手法と、H-NS 結合強度にもっとも寄与している配列特徴を特定するための手法を開発した。構築した回帰モデルの機械学習結果から、H-NS 結合強度、すなわちそれぞれの H-NS 結合領域の安定性を高めるためには、これまでに提案されてきた AT 含量、A-tract 配列というよりもむしろ、A あるいは T が局所的に連続する W5 連続配列が密集していることが条件となっている可能性を示唆した。このような柔軟な配列要素によって H-NS の結合が規定されることによって、第 2 章で得られた結果と第 3 章で得られた結果が矛盾なく共存し、H-NS 結合領域上で配列の多様性を高めながらも、H-NS の結合を維持し続けることが可能となるのだと考えられる。

5.1.2 H-NS による転写抑制の進化的意義

以上の結果から、H-NS の結合による利点は、外来性遺伝子の取り込み初期にそれらの発現を抑制して宿主に対する有害な効果を抑え込むことのみにあるのではなく、外来性遺伝子を長期的に保持し、転写抑制を介してそれらに蓄積する変異によって生じる有害な効果を抑えながら、それらが「仮想的に中立に」進化することを促し、適切に宿主のシステムに組み込まれるための環境を整えることにあると考えられる。

H-NS 結合領域下で多様性が高い遺伝子は、宿主のシステムに適応しているのではなく、単に遺伝子としての機能が壊れた偽遺伝子となっている可能性はあるか？第 3 章の解析結果から、その可能性は否定できると考える。偽遺伝子は、かつて遺伝子機能を担っていた DNA 配列が機能を失い、自由に変異が蓄積するようになったゲノム領域のことである。そのような領域では非同義座位、同義座位といった遺伝子機能に関わる DNA 配列の意味的区分が失われ、どちらも同じ程度の進化速度で変異を蓄積する (Kimura 1984)。H-NS 結合遺伝子では、H-NS 非結合遺伝子と比較して非同義座位多様性が有意に上昇しているとはいえ、それらは同義座位の多様性に比べて小さな値の中での差であった（図 3.5)。したがって、H-NS 結

合遺伝子であっても、それらの非同義座位は完全に中立的に進化しているわけではなく、自然選択が作用していると思われる。実際、大腸菌において H-NS 結合遺伝子のすべてがいっさい発現していないのではなく、通常の培養条件であっても、いくつかの H-NS 結合遺伝子の発現は見られる（図 3.12）。しかも、H-NS 結合の影響下にありながら発現が上昇している遺伝子の多くは、大腸菌系統群の分岐のはるか以前に獲得したと推測される「太古の外来性遺伝子」であった。以上の結果からわかるのは、H-NS の結合は、外来性遺伝子の発現を抑え込むことによって外来性遺伝子の偽遺伝子化を導くのではなく、長期的な保持とその間の進化の結果、外来性遺伝子の宿主システムへの適応という結果を導くということである。

H-NS による転写抑制効果の影響を受けている外来性遺伝子はなぜ、大腸菌のゲノムから失われないのか？生存にとって有害な遺伝子がゲノムから失われる詳細な分子生物学的メカニズムは明らかになっていないが、余計な遺伝子を排除する選択圧は非常に強力で、わずかにでも適応的コストのかかる遺伝子は急速に失われていくという例が、サルモネラ菌において報告されている (Koskiniemi et al. 2012)。一方、同様にサルモネラ菌において、*hns* 遺伝子を欠失させた変異体を用いた進化実験によって、野性型ではゲノム中に安定的に保持されているサルモネラ菌の III 型分泌装置に関連した遺伝子が、*hns* 変異体では急速に失われていくことが報告された (Ali et al. 2014)。したがって、H-NS の役割はむしろ、H-NS が存在しない場合には失われてしまう外来性遺伝子を、ゲノム中に保持し続けることにあると考えられる。同様の役割は、H-NS のパラログである Sfh タンパク質において知られており、保持に適応的コストがかかるために通常はすぐに排除されてしまうプラスミドが、そのプラスミド上にコードされた Sfh による“ステルス効果”によって菌体内に保持され得るという例が報告されている (Doyle et al. 2007)。外来性遺伝子は H-NS による転写抑制を受けているために、ゲノムから失われることなく適応していくのである。

HGT による外来性遺伝子の獲得は、もしそれが単に、ある細菌の何らかの形質を別の細菌が獲得することのみを意味するのであれば、必ずしも細菌系統群の形質の多様化や様々な環境への適応を導くメカニズムとはなり得ない。HGT が細菌の膨大な多様性を生み出しているのであれば、HGT による遺伝子交換に加えて、それぞれの細菌がその遺伝子をいつ使うか、いかに使うかといった、遺伝子制御の多様性を生み出す進化メカニズムが同時に存在しなければならない。大腸菌にお

いては、「遺伝子発現の多様性」が、大腸菌株間の表現型の多様性に大きく寄与していることが知られている。たとえば、コロニー形成に関わる Mat 線毛に関連した遺伝子群は大腸菌系統群で広く保存されているが、A 群大腸菌および B1 群大腸菌では通常発現しないような温度、pH などの環境条件において、B2 群大腸菌では Mat 関連遺伝子群を発現する (Lehti et al. 2013)。このような Mat 関連遺伝子の発現制御の多様性は、遺伝子のプロモータ領域に生じた配列多様性によってもたらされており、しかも Mat 遺伝子群のプロモータ領域は H-NS に覆われ、H-NS による発現抑制を受けているのである (Lehti et al. 2013)。このような遺伝子発現の多様性は、HGT によって獲得される遺伝子セットの多様性とともに、細菌がそれぞれの環境に適応する上で非常に重要である。H-NS は、大腸菌が様々な生息環境に適応する過程において、外来性遺伝子の獲得と、それら遺伝子の発現を最適化するプロセスに大きく関わってきたと考えられる。

大腸菌の進化における H-NS の役割を、最適化問題の解法とのアナロジーで考えることも可能である。最適化問題とは、ある集合上で定義された目的関数が最大値をとる元を求める問題である。局所最適解が大域的最適解と一致している場合は、各点における勾配ベクトルが求められれば最適化は簡単である。だが目的関数が多峰性の場合は（おそらく遺伝子型を定義域とした表現型の「適応度関数」は非常に強い多峰性である (Kauffman and Weinberger 1989)）、最適化は簡単ではない。そのような非線形・非凸・微分不可能な目的関数の最適化にはしばしば遺伝的アルゴリズムをはじめとした進化計算が使われる。進化計算では、多数のパーティクルを目的関数の定義域にばらまき、それぞれのパーティクルが個別に関数値や解の実行可能性を「評価」して、その評価に応じてパーティクルの自然選択が行われ、次の世代では生き残ったパーティクルの周辺をより重点的に探索する、というプロセスをとる。進化計算では、集団サイズや、それらの「多様性」をいかに制御するかが、アルゴリズムの性能を大きく左右する。多峰性の目的関数における最適解探索では、しばしば集団が特定の局所最適解周辺に縮退してしまう。それを避けるためには、各ステップで集団の多様性が維持できているか、すなわち関数定義域の広い範囲を探索できているかが重要となる (Kita and Yamamura 1999, Kobayashi 2009)。それぞれのパーティクルの「評価」をすべて厳密に取り扱い、低い評価値のパーティクルをすべて排除してしまうと、大域的最適解にたどり着く前に局所最適解に収束してしまう。したがって、多峰性の目的関数の最

適解探索では、ある程度の期間、それぞれのパーティクルの評価を保留し、十分に多様性を確保することが重要となる。これと同じ状況が、大腸菌の外来性遺伝子発現の最適化においてもおそらく生じている。外来性遺伝子の発現を生息環境に応じて「大域的最適化」する過程で、そこに至るまでの経路に存在する遺伝子型には、生存にとって有害な影響をもたらすものが無数にあると思われる。したがって、外来性遺伝子を取り込んですぐにその遺伝子を「評価」する、あるいは取り込み後の初期段階で評価してしまうと、そのような有害な遺伝子を持った個体は死滅し、HGT による多様性獲得の機会は失われる。一方、H-NS による転写抑制によって外来性遺伝子の評価を保留するならば、その外来性遺伝子を持った個体が集団中で増殖し、それぞれで配列に変異を蓄積して、多様性が十分に高まった状態で、外来性遺伝子の評価が可能となる。HGT の頻度はそれほど高くなく、世代あたり遺伝子あたり 10^{-6} から 10^{-4} と考えられており (Niehus et al. 2015)、すぐに有害な外来性遺伝子を排除してしまっては、その最適解探索を試みることもなく、多様性獲得の機会が失われてしまう。H-NS による外来性遺伝子の転写抑制は、最適解を探索する時間を稼ぐための重要な意義があると考えられる。また、遺伝子型の最適化問題解法としての H-NS の役割に関しては、焼き鈍し法（Simulated annealing）とのアナロジーも見いだされる。焼き鈍し法では、探索空間内の多様性を十分に確保したあと、徐々にパーティクルの評価と選別を厳しくし、探索領域を狭めていく。H-NS による転写抑制は、温度や浸透圧の変化などの環境ストレスによって解除されることが知られており (Prosseda et al. 2004)、ある程度集団中の配列多様性が大きくなった後に、外来性遺伝子がそれぞれの生息環境で評価されることで適応を遂げているのかもしれない。

高等生物においては、多様性はおもに遺伝子重複によって獲得されてきたと考えられている (Ohno 1970, Taylor and Raes 2004)。同じ機能を持つ遺伝子が２つ存在することで、一方の遺伝子の正常な機能を維持しながら、もう一方の遺伝子に自由な変異を蓄積することが可能となる。すなわち、重複遺伝子において選択圧が緩和される。これにより新たな遺伝子機能の獲得など、多様性が獲得される可能性を高めることができる。一方、細菌においてはタンパク質ファミリーの拡大に遺伝子重複がほとんど寄与してこなかったことが明らかにされている (Treangen et al. 2011)。細菌系統群はいかにして、現在地球上で観察される莫大な多様性を獲得するに至ったのか？本研究の結果から、少なくとも大腸菌においては、HGT

による外来性遺伝子の獲得と、H-NS による転写抑制を介した外来性遺伝子進化の選択圧の緩和が、多様性の獲得に大きな役割を担ってきたと考えられる。

5.2 今後の研究課題

最後に、本研究で未解決の問題、本研究からさらに発展させていくべき問題に関して、今後の研究課題として触れる。

H-NS 結合領域配列多様化の直接的なメカニズムはなにか？

本研究では、H-NS 結合領域の DNA 配列が、非結合領域の配列と比較して多様性が高い傾向にあることを明らかにしたが、それが実際にどのような生物学的メカニズムでなされているのか、十分に明らかにはなっていない。本研究では H-NS による転写抑制の間接的な効果として外来性遺伝子の選択圧が緩和され、進化速度が（見かけ上）増大する、という解釈をとった。この解釈は、同義座位および制御領域の存在しない遺伝子間領域において、H-NS の結合の有無によって多様性にほとんど差が見られないことを根拠としている。だが、何らかの未知のメカニズムによって、H-NS がその結合領域下の DNA 配列に直接的に変異を導入している可能性は排除しきれていない。H-NS 結合領域下の進化の加速が、H-NS の結合による直接的な効果によって生じているのか、転写抑制による間接的な効果によって生じているのか明らかにするためには、*in vitro* の進化実験、特に、野性型と *hns* 変異体の進化を実験室内進化によって比較することが必要である。*hns* 変異体の進化実験はサルモネラ菌などにおいて行われているが (Ali et al. 2014)、先行研究では H-NS 結合領域と非結合領域で多型の生じ方にどのような差があるかの比較検討は行っていない。H-NS 結合・非結合領域の変異頻度を野性型と *hns* 変異体で比較することによって、H-NS が進化の加速に直接的な影響を及ぼしているのかどうかを明らかにすることができると考えられる。

H-NS 非結合遺伝子で多様性が増大していた原因はなにか？

本研究では、外来性オーソログに関して、H-NS が結合していない場合にも多様性が大きく増大していることが観察された（図 3.9）。H-NS 非結合外来性オーソ

ログは、H-NS 結合外来性オーソログと同程度に多様性が増大している。これらの遺伝子はなぜ、どのようにして大腸菌系統群内の大きな多様性を獲得したのか？遺伝子の発現量の比較では、H-NS 非結合外来性オーソログは、H-NS 結合外来性オーソログと同じく、発現量が低かった（図 3.12）。これらの遺伝子に関しては H-NS の結合が見られないが、H-NS と似たなんらかの転写抑制因子が結合することによって、H-NS の場合と同様のメカニズムで多様性が増大している可能性がある。この仮説を検証するためには、これらの H-NS 非結合外来性オーソログ遺伝子について、それらの転写を制御している転写調節因子を特定することが必要である。

HGT シミュレーションの拡張

これまで、HGT をモデル化したシミュレーション研究によって (Niehus et al. 2015)、集団をスイープするのが遺伝子かゲノムか、すなわち、有利な形質を持った外来性遺伝子が集団中を二次的に水平伝播することによってそのような形質が集団中に固定されるのか、あるいは外来性遺伝子を獲得した個体が増殖することによって形質が集団中に固定されるのかが議論されている (Cohan and Koeppel 2008, Wiedenbeck and Cohan 2011)。HGT の頻度は非常に低いので、集団をスイープするのは常にゲノムであるはずだが、Niehus らによるシミュレーション研究では、外界から集団への個体の移入・移出が特定の条件を満たすと、遺伝子によるスイープが生じることが報告されている (Niehus et al. 2015)。しかし HGT の効果を考慮した生態学シミュレーション研究の多くでは、外来性遺伝子が獲得と同時に宿主にとって有利な形質をもたらすことを前提としたモデル化をしており、外来性遺伝子獲得初期の有害な効果や、H-NS による転写抑制による効果を組み込んだモデルは存在していない。H-NS によって外来性遺伝子の転写が抑制されることで、外来性遺伝子を取り込んだゲノムによる集団のスイープが必ずしも生じないかもしれない。微生物群集レベルの進化・多様化に関して、HGT が果たして来た役割を明らかにするためのシミュレーション研究を遂行する上で、H-NS の効果を取り入れたシミュレーションモデルの構築は大きな意義を持つと考える。

どれほど普遍的な現象か？

ゲノムの多様性を増加させ、それぞれの種が様々な環境条件に適応する上で、H-NS のような機能を持つタンパク質の保持は非常に優れた手段となっていると考えられる。したがって、大腸菌以外の多くの細菌のゲノムにおいても、このような性質を持つ転写抑制タンパク質が存在し、ゲノム上に分散した AT-rich 領域など、なんらかの配列的な特徴を持った局所領域をランドマークとして結合して、種の多様性を高めていることが予想される。H-NS のようなタンパク質の存在は、外来性遺伝子を積極的に導入して種の多様性を高めている細菌における、普遍的な進化システムなのではないか。事実、結核菌の持つ Lsr2 という転写抑制タンパク質は H-NS と配列的な相同性はないものの、ゲノム平均と比較して高い AT 含量の領域に結合するなど、H-NS と似た性質を有し (Gordon et al. 2011)、結核菌の進化過程において同様の役割を果たして来た可能性がある。細菌は地球上で最も多様性の高い生物である。この爆発的な多様性を獲得するために、H-NS を持たないほかの多くの細菌においても、H-NS と似た機能を持つ未知の転写抑制タンパク質が進化に寄与してきたのかもしれない。この仮説を検証するためには、今後より多くの細菌種に対して比較解析を実行し、そのような転写抑制タンパク質を見いだしていくことが必要である。

参考文献

Ahmed, Niyaz, Ulrich Dobrindt, Jörg Hacker, and Seyed E Hasnain. 2008. Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention. *Nature reviews microbiology* 6:387–394.

Akashi, Hiroshi. 2003. Translational selection and yeast proteome evolution. *Genetics* 164:1291–1303.

Ali, Sabrina S, Jeremy Soo, Chitong Rao, Andrea S Leung, David Hon-Man Ngai, Alexander W Ensminger, and William Wiley Navarre. 2014. Silencing by H-NS potentiated the evolution of Salmonella. *PLoS Pathogens* .

Anantha, Ravi P, Annette L McVeigh, Lanfong H Lee, Mary K Agnew, Frederick J Cassels, Daniel A Scott, Thomas S Whittam, and Stephen J Savarino. 2004. Evolutionary and functional relationships of colonization factor antigen i and other class 5 adhesive fimbriae of enterotoxigenic Escherichia coli. *Infection and immunity* 72:7190–7201.

Arold, Stefan T, Paul G Leonard, Gary N Parkinson, and John E Ladbury. 2010. H-NS forms a superhelical protein scaffold for DNA condensation. *Proceedings of the National Academy of Sciences* 107:15728–15732.

Atlung, Tove, and Hanne Ingmer. 1997. H-NS: a modulator of environmentally regulated gene expression. *Molecular microbiology* 24:7–17.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution* 17:540–552.

Ceschini, Simona, Giulio Lupidi, Massimo Coletta, Cynthia L Pon, Evandro Fioretti, and Mauro Angeletti. 2000. Multimeric self-assembly equilibria involv-

ing the histone-like protein H-NS. A thermodynamic study. *Journal of Biological Chemistry* 275:729–734.

Chattopadhyay, Sujay, Veronika Tchesnokova, Annette McVeigh, Dagmara I Kisiela, Kathleen Dori, Armando Navarro, Evgeni V Sokurenko, and Stephen J Savarino. 2012. Adaptive evolution of class 5 fimbrial genes in enterotoxigenic Escherichia coli and its functional consequences. *Journal of Biological Chemistry* 287:6150–6158.

Cohan, Frederick M, and Alexander F Koeppel. 2008. The origins of ecological diversity in prokaryotes. *Current Biology* 18:R1024–R1034.

Crothers, Donald M, Tali E Haran, and James G Nadeau. 1990. Intrinsically bent DNA. *Journal of Biological Chemistry* 265:7093–7096.

Dame, Remus T. 2005. The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin. *Molecular microbiology* 56:858–870.

Dame, Remus T, Olga J Kalmykowa, and David C Grainger. 2011. Chromosomal macrodomains and associated proteins: implications for DNA organization and replication in gram negative bacteria. *PLoS genetics* 7:e1002123.

Dame, Remus T, Martijn S Luijsterburg, Evelyne Krin, Philippe N Bertin, Rolf Wagner, and Gijs JL Wuite. 2005. DNA bridging: a property shared among H-NS-like proteins. *Journal of bacteriology* 187:1845–1848.

Dame, Remus T, Maarten C Noom, and Gijs JL Wuite. 2006. Bacterial chromatin organization by H-NS protein unravelled using dual DNA manipulation. *Nature* 444:387–390.

Darling, Aaron CE, Bob Mau, Frederick R Blattner, and Nicole T Perna. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research* 14:1394–1403.

Darling, Aaron E, Bob Mau, and Nicole T Perna. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS one* 5:e11147.

Datsenko, Kirill A, and Barry L Wanner. 2000. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proceedings of the National Academy of Sciences* 97:6640–6645.

Datta, Naomi, and Polyxeni Kontomichalou. 1965. Penicillinase synthesis controlled by infectious R factors in Enterobacteriaceae. *Nature* 208:239.

Davids, Wagied, and Zhaolei Zhang. 2008. The impact of horizontal gene transfer in shaping operons and protein interaction networks–direct evidence of preferential attachment. *BMC evolutionary biology* 8:23.

Doolittle, W Ford. 2000. Uprooting the tree of life. *Scientific American* 282:90.

Dorman, Charles J. 2004. H-NS: a universal regulator for a dynamic genome. *Nature Reviews Microbiology* 2:391–400.

Dorman, Charles J. 2007. H-NS, the genome sentinel. *Nature Reviews Microbiology* 5:157–161.

Dorman, Charles J. 2013. Genome architecture and global gene regulation in bacteria: making progress towards a unified model? *Nature Reviews Microbiology* 11:349–355.

Doyle, Marie, Maria Fookes, AL Ivens, Michael W Mangan, John Wain, and Charles J Dorman. 2007. An H-NS-like stealth protein aids horizontal DNA transmission in bacteria. *Science* 315:251–252.

Drummond, D Allan, Jesse D Bloom, Christoph Adami, Claus O Wilke, and Frances H Arnold. 2005. Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences* 102:14338–14343.

Falkowski, Paul G, Tom Fenchel, and Edward F Delong. 2008. The microbial engines that drive Earth's biogeochemical cycles. *Science* 320:1034–1039.

Fang, Ferric C, and Sylvie Rimsky. 2008. New insights into transcriptional regulation by H-NS. *Current opinion in microbiology* 11:113–120.

Furuya, E Yoko, and Franklin D Lowy. 2006. Antimicrobial-resistant bacteria in the community setting. *Nature Reviews Microbiology* 4:36–45.

Gao, Feng, Hao Luo, and Chun-Ting Zhang. 2013. DoriC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes. *Nucleic acids research* 41:D90–D93.

Gao, Feng, and Chun-Ting Zhang. 2007. DoriC: a database of oriC regions in bacterial genomes. *Bioinformatics* 23:1866–1867.

Garcia-Vallve, Santiago, Eduard Guzmán, MA Montero, and Antoni Romeu. 2003. HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic acids research* 31:187–189.

Gordon, Blair RG, Yifei Li, Atina Cote, Matthew T Weirauch, Pengfei Ding, Timothy R Hughes, William Wiley Navarre, Bin Xia, and Jun Liu. 2011. Structural basis for recognition of AT-rich DNA by unrelated xenogeneic silencing proteins. *Proceedings of the National Academy of Sciences* 108:10690–10695.

Grainger, David C, Douglas Hurd, Martin D Goldberg, and Stephen JW Busby. 2006. Association of nucleoid proteins with coding and non-coding segments of the Escherichia coli genome. *Nucleic acids research* 34:4642–4652.

Guindon, Stephane, and Olivier Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology* 52:696–704.

Hao, Weilong, and G Brian Golding. 2006. The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Research* 16:636–643.

Hasegawa, Masami, Hirohisa Kishino, and Taka-aki Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution* 22:160–174.

Hatoum-Aslan, Asma, and Luciano A Marraffini. 2014. Impact of crispr immunity on the emergence and virulence of bacterial pathogens. *Current opinion in microbiology* 17:82–90.

Hayashi, Koji, Naoki Morooka, Yoshihiro Yamamoto, Katsutoshi Fujita, Katsumi Isono, Sunju Choi, Eiichi Ohtsubo, Tomoya Baba, Barry L Wanner, Hirotada Mori, et al. 2006. Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. *Molecular systems biology* 2.

Hayashi, Tetsuya, Kozo Makino, Makoto Ohnishi, Ken Kurokawa, Kazuo Ishii, Katsushi Yokoyama, Chang-Gyun Han, Eiichi Ohtsubo, Keisuke Nakayama, Takahiro Murata, et al. 2001. Complete genome sequence of enterohemorrhagic Eschelichia coli O157: H7 and genomic comparison with a laboratory strain K-12. *DNA research* 8:11–22.

Holder, Mark T. 2015. Phylogenetic Methods. `http://phylo.bio.ku.edu/content/biol-848-phylogenetic-methods-10`.

Hommais, Florence, Evelyne Krin, Christine Laurent-Winter, Olga Soutourina, Alain Malpertuy, Jean-Pierre Le Caer, Antoine Danchin, and Philippe Bertin. 2001. Large-scale monitoring of pleiotropic regulation of gene expression by the prokaryotic nucleoid-associated protein, H-NS. *Molecular microbiology* 40:20–36.

Hooper, Sean D, and Otto G Berg. 2003. Duplication is more common among laterally transferred genes than among indigenous genes. *Genome Biol* 4:R48.

Horvath, Philippe, and Rodolphe Barrangou. 2010. CRISPR/Cas, the immune system of bacteria and archaea. *Science* 327:167–170.

Ishikawa, Shu, Yoshitoshi Ogura, Mika Yoshimura, Hajime Okumura, Eunha Cho, Yoshikazu Kawai, Ken Kurokawa, Taku Oshima, and Naotake Ogasawara. 2007. Distribution of stable DnaA-binding sites on the Bacillus subtilis genome detected using a modified ChIP-chip method. *DNA research* 14:155–168.

Jannasch, Holger W, and Michael J Mottl. 1985. Geomicrobiology of deep-sea hydrothermal vents. *Science* 229:717–725.

Kahramanoglou, Christina, Aswin SN Seshasayee, Ana I Prieto, David Ibberson, Sabine Schmidt, Jurgen Zimmermann, Vladimir Benes, Gillian M Fraser, and

Nicholas M Luscombe. 2011. Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli. *Nucleic acids research* 39:2073–2091.

Katoh, Kazutaka, and Hiroyuki Toh. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in bioinformatics* 9:286–298.

Kauffman, Stuart A, and Edward D Weinberger. 1989. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of theoretical biology* 141:211–245.

Kent, W James. 2002. BLAT—the BLAST-like alignment tool. *Genome research* 12:656–664.

Kimura, Motoo. 1984. *The neutral theory of molecular evolution*. Cambridge University Press.

Kita, Hajime, and Masayuki Yamamura. 1999. Design Guidelines for Genetic Algorithms based on Function Specialization Hypothesis. *Journal of The Society of Instrument and Control Engineers* 38:612–617.

Kobayashi, Shigenobu. 2009. The Frontiers of Real-coded Genetic Algorithms. *Transactions of the Japanese Society for Artificial Intelligence* 24:147–162.

Koonin, Eugene V, and Yuri I Wolf. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic acids research* 36:6688–6719.

Koskiniemi, Sanna, Song Sun, Otto G Berg, and Dan I Andersson. 2012. Selection-driven gene loss in bacteria. *PLoS Genet* 8:e1002787–e1002787.

Kotlajich, Matthew V, Daniel R Hron, Beth A Boudreau, Zhiqiang Sun, Yuri L Lyubchenko, and Robert Landick. 2015. Bridged filaments of histone-like nucleoid structuring protein pause rna polymerase and aid termination in bacteria. *Elife* 4:e04970.

Ku, Chuan, Shijulal Nelson-Sathi, Mayo Roettger, Filipa L Sousa, Peter J Lockhart, David Bryant, Einat Hazkani-Covo, James O McInerney, Giddy Landan,

and William F Martin. 2015. Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* 524:427–432.

Kuraku, Shigehiro. 2013. Impact of asymmetric gene repertoire between cyclostomes and gnathostomes. In *Seminars in cell & developmental biology*, volume 24, 119–127. Elsevier.

Lang, Benjamin, Nicolas Blot, Emeline Bouffartigues, Malcolm Buckle, Marcel Geertz, Claudio O Gualerzi, Ramesh Mavathur, Georgi Muskhelishvili, Cynthia L Pon, Sylvie Rimsky, et al. 2007. High-affinity DNA binding sites for H-NS provide a molecular basis for selective silencing within proteobacterial genomes. *Nucleic acids research* 35:6330–6337.

Langmead, Ben, and Steven L Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* 9:357–359.

Lawrence, Jeffrey G, and Howard Ochman. 1997. Amelioration of bacterial genomes: rates of change and exchange. *Journal of molecular evolution* 44:383–397.

Lawrence, Jeffrey G, and Howard Ochman. 1998. Molecular archaeology of the Escherichia coli genome. *Proceedings of the National Academy of Sciences* 95:9413–9417.

Lee, David J, Lewis EH Bingle, Karin Heurlier, Mark J Pallen, Charles W Penn, Stephen JW Busby, and Jon L Hobman. 2009. Gene doctoring: a method for recombineering in laboratory and pathogenic Escherichia coli strains. *BMC microbiology* 9:252.

Lehti, Timo A, Philippe Bauchart, Maini Kukkonen, Ulrich Dobrindt, Timo K Korhonen, and Benita Westerlund-Wikström. 2013. Phylogenetic group-associated differences in regulation of the common colonization factor Mat fimbria in Escherichia coli. *Molecular microbiology* 87:1200–1222.

Leplae, Raphaël, Aline Hebrant, Shoshana J Wodak, and Ariane Toussaint. 2004.

ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic acids research* 32:D45–D49.

Lercher, Martin J, and Csaba Pál. 2008. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Molecular biology and evolution* 25:559–567.

Levin, Bruce R, and Carl T Bergstrom. 2000. Bacteria are different: observations, interpretations, speculations, and opinions about the mechanisms of adaptive evolution in prokaryotes. *Proceedings of the National Academy of Sciences* 97:6981–6985.

Levin, Bruce R, Véronique Perrot, and Nina Walker. 2000. Compensatory mutations, antibiotic resistance and the population genetics of adaptive evolution in bacteria. *Genetics* 154:985–997.

Ley, Ruth E, Daniel A Peterson, and Jeffrey I Gordon. 2006. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* 124:837–848.

Li, Li, Christian J Stoeckert, and David S Roos. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* 13:2178–2189.

Lucchini, Sacha, Gary Rowley, Martin D Goldberg, Douglas Hurd, Marcus Harrison, and Jay CD Hinton. 2006. H-NS mediates the silencing of laterally acquired genes in bacteria. *PLoS Pathogens* 2:e81.

Lunter, Gerton, Andrea Rocco, Naila Mimouni, Andreas Heger, Alexandre Caldeira, and Jotun Hein. 2008. Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome research* 18:298–309.

Maurer, Sebastian, Jürgen Fritz, and Georgi Muskhelishvili. 2009. A systematic In Vitro study of nucleoprotein complexes formed by bacterial nucleoid-associated proteins revealing novel types of DNA organization. *Journal of molecular biology* 387:1261–1276.

McFall-Ngai, Margaret, Michael G Hadfield, Thomas CG Bosch, Hannah V Carey, Tomislav Domazet-Lošo, Angela E Douglas, Nicole Dubilier, Gerard Eberl, Tadashi Fukami, Scott F Gilbert, et al. 2013. Animals in a bacterial world, a new imperative for the life sciences. *Proceedings of the National Academy of Sciences* 110:3229–3236.

Molina, Nacho, and Erik van Nimwegen. 2008. Universal patterns of purifying selection at noncoding positions in bacteria. *Genome research* 18:148–160.

Nakamura, Yoji, Takeshi Itoh, Hideo Matsuda, and Takashi Gojobori. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature genetics* 36:760–766.

Navarre, William Wiley, Michael McClelland, Stephen J Libby, and Ferric C Fang. 2007. Silencing of xenogeneic DNA by H-NS—facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA. *Genes & Development* 21:1456–1471.

Navarre, William Wiley, Steffen Porwollik, Yipeng Wang, Michael McClelland, Henry Rosen, Stephen J Libby, and Ferric C Fang. 2006. Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella. *Science* 313:236–238.

Nelson-Sathi, Shijulal, Filipa L Sousa, Mayo Roettger, Nabor Lozada-Chávez, Thorsten Thiergart, Arnold Janssen, David Bryant, Giddy Landan, Peter Schönheit, Bettina Siebers, et al. 2014. Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* .

Niehus, Rene, Sara Mitri, Alexander G Fletcher, and Kevin R Foster. 2015. Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nature communications* 6.

Ochman, Howard, Jeffrey G Lawrence, and Eduardo A Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299–304.

Ogura, Yoshitoshi, Tadasuke Ooka, Atsushi Iguchi, Hidehiro Toh, Md Asadulghani, Kenshiro Oshima, Toshio Kodama, Hiroyuki Abe, Keisuke Nakayama, Ken Kurokawa, et al. 2009. Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic Escherichia coli. *Proceedings of the National Academy of Sciences* 106:17939–17944.

Ohno, Susumu. 1970. *Evolution by gene duplication*. Springer Science & Business Media.

Oshima, Kenshiro, Hidehiro Toh, Yoshitoshi Ogura, Hiroyuki Sasamoto, Hidetoshi Morita, Sang-Hee Park, Tadasuke Ooka, Sunao Iyoda, Todd D Taylor, Tetsuya Hayashi, et al. 2008. Complete genome sequence and comparative analysis of the wild-type commensal Escherichia coli strain SE11 isolated from a healthy adult. *DNA research* 15:375–386.

Oshima, Taku, Shu Ishikawa, Ken Kurokawa, Hirofumi Aiba, and Naotake Ogasawara. 2006. Escherichia coli histone-like protein H-NS preferentially binds to horizontally acquired DNA in association with RNA polymerase. *DNA research* 13:141–153.

Park, Peter J. 2009. Chip-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* 10:669–680.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research* 12:2825–2830.

Peters, Jason M, Abbey D Vangeloff, and Robert Landick. 2011. Bacterial transcription terminators: the RNA 3′ -end chronicles. *Journal of molecular biology* 412:793–813.

Prosseda, Gianni, Maurizio Falconi, Mara Giangrossi, Claudio O Gualerzi, Gioacchino Micheli, and Bianca Colonna. 2004. The virF promoter in Shigella: more than just a curved DNA stretch. *Molecular microbiology* 51:523–537.

Rimsky, Sylvie. 2004. Structure of the histone-like protein H-NS and its role in regulation and genome superstructure. *Current opinion in microbiology* 7:109–114.

Rocha, Eduardo PC. 2004. The replication-related organization of bacterial genomes. *Microbiology* 150:1609–1627.

Salgado, Heladia, Martin Peralta-Gil, Socorro Gama-Castro, Alberto Santos-Zavaleta, Luis Muñiz-Rascado, Jair S García-Sotelo, Verena Weiss, Hilda Solano-Lira, Irma Martínez-Flores, Alejandra Medina-Rivera, et al. 2013. RegulonDB v8. 0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic acids research* 41:D203–D213.

Schwarz, Gideon, et al. 1978. Estimating the dimension of a model. *The annals of statistics* 6:461–464.

Sette, Marco, Roberto Spurio, Edoardo Trotta, Cinzia Brandizi, Anna Brandi, Cynthia L Pon, Gaetano Barbato, Rolf Boelens, and Claudio O Gualerzi. 2009. Sequence-specific recognition of DNA by the C-terminal domain of nucleoid-associated protein H-NS. *Journal of Biological Chemistry* 284:30453–30462.

Sharp, Paul M. 1991. Determinants of DNA sequence divergence between Escherichia coli and Salmonella typhimurium: Codon usage, map position, and concerted evolution. *Journal of molecular evolution* 33:23–33.

Sharp, Paul M, Elizabeth Cowe, Desmond G Higgins, Denis C Shields, Kenneth H Wolfe, and Frank Wright. 1988. Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity. *Nucleic acids research* 16:8207–8211.

Siguier, Patricia, Jocelyne Pérochon, L Lestrade, Jacques Mahillon, and Michael Chandler. 2006. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic acids research* 34:D32–D36.

Smyth, Clare P, Thomas Lundbäck, Debora Renzoni, Giuliano Siligardi, Rebecca Beavil, Meredith Layton, Julie M Sidebotham, Jay CD Hinton, Paul C Driscoll, Christopher F Higgins, et al. 2000. Oligomerization of the chromatin-structuring protein H-NS. *Molecular microbiology* 36:962–972.

Spurio, Roberto, Maurizio Falconi, Anna Brandi, Cynthia L Pon, and Claudio O Gualerzi. 1997. The oligomeric structure of nucleoid protein H-NS is necessary for recognition of intrinsically curved DNA and for DNA bending. *The EMBO journal* 16:1795–1805.

Stoebel, Daniel M, Andrew Free, and Charles J Dorman. 2008. Anti-silencing: overcoming H-NS-mediated repression of transcription in Gram-negative enteric bacteria. *Microbiology* 154:2533–2545.

Taniguchi, Yuichi, Paul J Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X Sunney Xie. 2010. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329:533–538.

Tavaré, Simon. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences* 17:57–86.

Taylor, John S, and Jeroen Raes. 2004. Duplication and divergence: the evolution of new genes and old ideas. *Annu. Rev. Genet.* 38:615–643.

Tobe, Toru, Masanosuke Yoshikawa, and Chihiro Sasakawa. 1995. Thermoregulation of virB transcription in Shigella flexneri by sensing of changes in local DNA superhelicity. *Journal of bacteriology* 177:1094–1097.

Toh, Hidehiro, Kenshiro Oshima, Atsushi Toyoda, Yoshitoshi Ogura, Tadasuke Ooka, Hiroyuki Sasamoto, Sang-Hee Park, Sunao Iyoda, Ken Kurokawa, Hidetoshi Morita, et al. 2010. Complete genome sequence of the wild-type commensal Escherichia coli strain SE15, belonging to phylogenetic group B2. *Journal of bacteriology* 192:1165–1166.

Torsvik, Vigdis, Jostein Goksøyr, and Frida Lise Daae. 1990. High diversity in DNA of soil bacteria. *Applied and environmental microbiology* 56:782–787.

Touchon, Marie, Claire Hoede, Olivier Tenaillon, Valérie Barbe, Simon Baeriswyl, Philippe Bidet, Edouard Bingen, Stéphane Bonacorsi, Christiane Bouchier, Odile Bouvet, et al. 2009. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet* 5:e1000344.

Treangen, Todd J, Eduardo P Rocha, et al. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* 7:e1001284.

Tsoy, Olga V, Mikhail A Pyatnitskiy, Marat D Kazanov, and Mikhail S Gelfand. 2012. Evolution of transcriptional regulation in closely related bacteria. *BMC evolutionary biology* 12:200.

Uchiyama, Ikuo. 2003. MBGD: microbial genome database for comparative analysis. *Nucleic acids research* 31:58–62.

Uchiyama, Ikuo. 2007. MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. *Nucleic acids research* 35:D343–D346.

Uchiyama, Ikuo, Toshio Higuchi, and Mikihiko Kawai. 2010. MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. *Nucleic acids research* 38:D361–D365.

Uchiyama, Ikuo, Motohiro Mihara, Hiroyo Nishide, and Hirokazu Chiba. 2015. MBGD update 2015: microbial genome database for flexible ortholog analysis utilizing a diverse set of genomic data. *Nucleic acids research* 43:D270–D276.

Ueda, Takeshi, Hiroki Takahashi, Ebru Uyar, Shu Ishikawa, Naotake Ogasawara, and Taku Oshima. 2013. Functions of the Hha and YdgT proteins in transcriptional silencing by the nucleoid proteins, H-NS and StpA, in Escherichia coli. *DNA research* 20:263–271.

Ueguchi, Chiharu, Chiharu Seto, Tomomi Suzuki, and Takeshi Mizuno. 1997. Clarification of the dimerization domain and its functional significance for the Escherichia coli nucleoid protein H-NS. *Journal of molecular biology* 274:145–151.

Uyar, Ebru, Ken Kurokawa, Mika Yoshimura, Shu Ishikawa, Naotake Ogasawara, and Taku Oshima. 2009. Differential binding profiles of StpA in wild-type and hns mutant cells: a comparative analysis of cooperative partners by chromatin immunoprecipitation-microarray analysis. *Journal of bacteriology* 191:2388–2391.

Wade, Joseph T, and David C Grainger. 2014. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nature Reviews Microbiology* 12:647–653.

Wang, Wenqin, Gene-Wei Li, Chongyi Chen, X Sunney Xie, and Xiaowei Zhuang. 2011. Chromosome organization by a nucleoid-associated protein in live bacteria. *Science* 333:1445–1449.

Wiedenbeck, Jane, and Frederick M Cohan. 2011. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS microbiology reviews* 35:957–976.

Williamson, Helen S, and Andrew Free. 2005. A truncated H-NS-like protein from enteropathogenic Escherichia coli acts as an H-NS antagonist. *Molecular microbiology* 55:808–827.

Yang, Ziheng. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences: CABIOS* 13:555–556.

Yang, Ziheng. 2006. *Computational molecular evolution*, volume 21. Oxford University Press Oxford.

Zuber, F, D Kotlarz, S Rimsky, and H Buc. 1994. Modulated expression of promoters containing upstream curved DNA sequences by the Escherichia coli nucleoid protein H-NS. *Molecular microbiology* 12:231–240.

謝辞

本研究を進めるにあたり、指導教員として終始暖かいご指導を賜り、援助してくださった東京工業大学　黒川　顕　教授に心より厚く御礼を申し上げます。

日頃から多岐にわたるご指導をいただきました東京工業大学　山田　拓司　講師、森　宙史　助教に深く感謝いたします。

本研究の実験を行ってくださった奈良先端科学技術大学院大学の小笠原　直毅　教授、大島　拓　助教、石川　周　助教、Ebru Uyar 博士に心から感謝申し上げます。特に大島　拓　助教には、本研究の様々な局面において多くのアドバイスをいただきました。厚く御礼を申し上げます。本研究の実験系を構築してくださり、ご指導、ご助言をいただきました大阪大学　戸邉　亨　教授に深く御礼を申し上げます。

ChAP-seq サンプルのシーケンシングを行っていただいた東京大学の鈴木　穣　教授、広島大学の金井　昭教　助教に心から感謝いたします。

博士課程在学中は、東京工業大学博士課程教育リーディングプログラム 情報生命博士教育院より経済的なご支援をいただき、教育プログラムを通して多くの知的刺激をいただきました。厚く御礼申し上げます。

東京工業大学　黒川研究室の皆様には、様々な面でお世話になりました。心から感謝いたします。

最後に、長きにわたる学生生活を様々な面で支え続けてくれた両親と妹に心から感謝いたします。