UNIVERSITÉ FRANÇOIS RABELAIS DE TOURS

École Doctorale Santé, Sciences, Technologies

LABORATOIRE DE MATHÉMATIQUES ET PHYSIQUE THÉORIQUE

THÈSE présenté par :

Shuangwei HU

soutenue le 01 decembre 2011

pour obtenir le grade de : Docteur de l'Université François - Rabelais de Tours

Discipline/ Spécialité : Physique Théorique

APPLICATION DE LA SYMETRIE DE JAUGE ET DE LA THEORIE DES SOLITONS AUX PROTEINES REPLIEES

THÈSE DIRIGÉE PAR :

M. NIEMI Antti	Directeur de Recherche CNRS, Université de Tours, France

RAPPORTEURS :

M. KNELLER Gerald	Professeur, Université d'Orléans, France
M. PLOTKIN Steven Samuel	Professeur, Université de la Colombie-Britannique, Canada

JURY :

M. BACHAS Costas	Directeur de Recherche CNRS, Ecole Normale Supérieure (Paris), France
M. GAREL Thomas	Directeur de Recherche CNRS, IPhT Saclay, France
M. GIACOMINI Hector	Professeur, Université de Tours, France
M. KNELLER Gerald	Professeur, Université d'Orléans, France
MME LESIEUR Claire	Chargé de Recherche CNRS, Université de Savoie, France
M. NIEMI Antti	Directeur de Recherche CNRS, Université de Tours, France
M. PLOTKIN Steven Samuel	Professeur, Université de la Colombie-Britannique, Canada

Remerciements

Je dois beaucoup à mon directeur Antti J. Niemi, beaucoup plus que je ne peux le remercier par les mots. J'ai été avec bonheur porté par lui dans le monde fantastique de la recherche sur les protéines, dont l'exigence coïncide avec mon expérience précédente des polymères et de la physique. Ses idées et encouragements ont permis des percées lorsque j'étais confronté à certains problèmes. Par exemple, il a affirmé qu'il s'agit vraisemblablement d'un canard, si vous entendez cancaner. Je n'oublierai jamais ces moments où nous avons attrapé un canard, surtout pour la première fois, où nous avons découvert le visage mystérieux du soliton en liaison avec des protéines. Le présent travail a témoigné de ses conseils considérable sur l'idée principale, l'organisation et l'écriture. Je le remercie aussi pour son amitié, sa gentillesse et son soutien.

Je remercie Gerald Kneller d'avoir accepté d'être rapporteur de ma thèse, en particulier le remercier pour sa présence. Steven Samuel Plotkin a également accepté la lourde tâche d'écrire l'autre rapport de ma thèse. Je suis particulièrement reconnaissant envers Thomas Garel et Claire Lesieur pour leur lecture détaillée du manuscrit et leurs nombreuses remarques pertinentes. Merci également à Costas Bachas Hector Giacomini pour avoir relu mon manuscrit et leurs encouragements.

Un grand merci aussi à tous les collègues du LMPT. Je suis reconnaissant envers Maxim Chernodub et Stam Nicolis pour nos différentes discussions qui m'ont inspirées, et spécialement à Stam pour avoir eu une vue d'ensemble sur mon projet de thèse. Je tiens à remercier encore une fois Hector Giacomini pour son aide précieuse intervenue à de nombreux moments, qu'importe les questions académiques ou les problèmes d'administration. Je remercie également Emmanuel Lesigne et Laurent Véron pour leur hospitalité et leur aide sur les questions d'administration. J'ai également été heureux et chanceux de profiter de bons moments avec Jérémy Le Deunff, Julien Garaud, Jean-Paul Ngome, Francesco Sardelli, Elisa Meunier, Kévin Morand, Anh Dao Nguyen, Safaa El Sayed, Rim Essifi, Lingmin Liao, Ali Makki, Tai Nguyen Phuoc, Julie Oger, Silvia Sastre, Yinna Ye, Rami Younes, et bien d'autres. Votre soutien au cours des dernières années restera un grand souvenir.

Je suis également redevable aux gens de la division de physique théorique du département de physique et d'astronomie de l'Université d'Uppsala en Suède. La moitié de mes études de doctorat là-bas et plus tard mes visites m'ont permis de m'améliorer en physique et en biologie. Merci Ulf Lindström pour son hospitalité et son soutien. Des remerciements

particuliers doivent être consacrés à Martin Lundgren, Andrey Krokhotin, Xubiao Peng et Yan Kou. La thèse ne serait pas complète sans les discussions utiles que j'ai eu avec vous tous.

Ma profonde gratitude va à mon maître de thèse Molin Ge qui m'a guidé à la beauté de la physique. Sa confiance en moi a toujours été un grand encouragement. Aussi, je n'aurais pas été en mesure de profiter de la recherche juste après mon arrivée ici, sans l'éducation reçue à la fois à l'Université de Tianjin et à l'Université de Nankai en Chine.

Je suis reconnaissant envers mes amis mais je crains qu'il soit difficile de donner une liste complète, je viens de vérifier mon compte Facebook. Ici, je voudrais remercier tout particulièrement Bruno-Marie Dupin, mon entraîneur de voile, pour sa formation enthousiaste et aussi remercier mon coéquipier Qian Zhou pour sa patience et son sens social. L'Open de France à Quiberon restera un souvenir inoubliable dans ma vie. J'ai aussi manquer le bon moment avec mes amis Fang Luo, Wangshu Jiang, Cheng Zhong, Wenxian Dong, Jane Xu ...

Je remercie ma famille pour leur amour et leur confiance. Je suis très reconnaissant surtout envers mes parents, Juguang Hu and Shuxia Zhang, qui ont vécu des moments difficiles à construire ma génération. Je remercie ma sœur aînée Cuijuan Hu et mon frère Zhengwei Hu de m'avoir soutenu par quelques solutions raisonnables quand j'en ai eu besoin. Je dois beaucoup à ma famille. Merci ma petite amie Qian Dong pour son amour et ses encouragements.

Je pourrais continuer ces remerciements à l'infini, mais je dois m'arrêter pour une question de place. Merci à tous !

Résumé

La structure des protéines est organisée de manière hiérarchique, des séquences d'acides aminés aux hélices α et feuillets β réguliers au niveau secondaire et finalement les formes compactes tertiaires. Dans la cellule, la chaîne polypeptidique d'une protéine, récemment synthétisée, subit un effondrement spontané vers la structure tertiaire adéquate, dans le but d'accomplir sa fonction spécifique. Après plus de cinquante ans d'investigation scientifique sur ce problème classique de repliement des protéines une solution semble lentement se dessiner à l'horizon. Le but de cette thèse est d'étudier plus en profondeur le repliement des protéines, au moyen des concepts d'invariance de jauge et d'universalité.

La structure de jauge émerge de l'équation de Frenet qui est utilisée pour décrire la forme de la chaîne principale de la protéine. Le principe d'invariance de jauge en théorie quantique des champs conduit à une fonctionnelle d'énergie effective pour une protéine, développée dans le but d'extraire les propriétés universelles des protéines repliées durant la phase d'effondrement, et qui est caractérisée par la loi d'échelle du rayon de giration au niveau tertiaire de la structure protéique [11].

Dans cette thèse, l'existence d'une large universalité au niveau secondaire de la structure protéique est étudiée à l'aide de la théorie des solitons. La fonctionnelle d'énergie invariante de jauge alliée à l'équation de Frenet discrète conduit à une solution solitonique de type *kink*, identifiée comme un motif hélice-boucle-hélice dans la protéine. Les paramètres qui caractérisent un repliement particulier de protéine sont tous globaux au niveau secondaire, allant au-delà de tous les détails et complexités des acides aminés et de leurs interactions. Cette théorie des solitons peut être vue comme la manifestation évidente de l'observation expérimentale que le nombre de structures protéiques, dans la nature, est plus petit que le nombre de séquences d'acides aminés [10].

Le repliement de la chaîne principale de protéines entières est donc construit en assemblant plusieurs solitons. Nous présentons ici la relation étroite entre notre modèle et l'équation de Schrodinger non linéaire et l'utilisons afin d'accélérer les simulations numériques. la réussite de cette description des boucles plus longues et des boucles qui connectent les hélices α avec les feuillets β. La modélisation d'un nombre de protéines biologiquement actives reproduit la structure naturelle avec une précision expérimentale.

Ce travail présenté dans cette thèse pourrait ouvrir des portes vers de nouvelles voies d'exploration concernant le problème du repliement de protéines. Le modèle développé ici constitue une base solide pour commencer et pourra être facilement adapté afin d'inclure l'interaction spécifique entre les acides aminés et ainsi décrire la dynamique des protéines, dans le but final d'aborder des questions importantes tels que le mécanisme de repliement

correct et incorrect des protéines.

Mots clés : Le repliement des protéines, des équations de Frenet, symétrie de jauge, soliton, l'universalité

Abstract

Protein structure is organized hierarchically from the primary amino acid sequence, to regular α-helices and β-strands at the secondary level, and finally to the tertiary compact shape. In the cell, the newly synthesized polypeptide chain of a protein undergoes a spontaneous collapse to the proper tertiary structure, in order to perform its specific function. After more than fifty years of scientific inquiry on this classic problem of protein folding, a solution is slowly rising on the horizon. The purpose of this thesis is to further investigate protein folding, by means of the general concepts of gauge invariance and universality.

The gauge structure emerges in the Frenet equation which is utilized to describe the shape of protein backbone. The gauge invariance principle in quantum field theory leads us an effective energy functional for a protein, which has been found to catch the universal properties of folded proteins in their collapse phase, characterized by the scaling law of gyration radius on the tertiary level of protein structure [11].

In this thesis, the existence of wide universality on the secondary level of protein structure is investigated, in terms of soliton theory. The synthesis of the gauge-invariant energy functional with the discrete Frenet equation leads to a kink soliton solution, which is identified as the helix-loop-helix motif in protein. The parameters that characterize a particular protein fold are all global on the secondary level, going beyond all the details and complexities of amino acids and their interactions. This soliton theory can be viewed as the obvious manifestation of experimental observation that the number of protein structures in nature is quite more limited than the number of amino acid sequences [10].

The main-chain folding of entire proteins is then built by assembling multiple solitons. We present the intimate connection between our model and a generalized nonlinear Schrodinger equation and use it to speed up the simulation. The flexibility of our approach ensures the successful description of longer loops and loops connecting α-helices with β-strands. The modeling of a number of biologically active proteins reproduces the native structure with experimental accuracy.

We believe that the the work in this thesis will open doors to new ways of the future research on the protein folding problem. The model we develop forms a solid basis to start and will be easily extended to include the sequence specific interaction so that it can describe the dynamics of proteins, with the goal to finally address issues such as the mechanism of protein folding and misfolding.

Keywords : Protein folding, Frenet equations, gauge symmetry, soliton, universality.

Table des matières

Première partie

Synopsis

Chapitre 1

Introduction générale

Les protéines sont impliquées dans presque toutes les fonctions cellulaires, comprenaient principalement la liaison de fixer une autre molécule, la catalyse, le rôle de commutateurs moléculaires et celui de composant structural [48]. Afin de réaliser une fonction donnée, la plupart des protéines ont besoin de se replier en une structure tridimensionnelle unique, appelé état natif [43]. Un repliement incorrect peut avoir des conséquences désastreuses comme il a été démontré par l'existence des prions ou encore des agrégats protéiques intervenant dans la maladie d'Alzheimer. Par conséquent, la détermination de la structure joue un rôle central dans la recherche des protéines et a de nombreuses applications dans le domaine de la biologie ou médicine.

Alors que la chaîne polypeptidique d'une protéine est synthétisée en utilisant l'information de l'ARN messager comme un guide, l'étape suivante qui conduit à sa conformation naturelle, s'effectue, en réalité, de manière spontanée. Il a été postulé par Anfinsen et ses collaborateurs dans les années cinquante que l'état naturelle d'une protéine correspond au minimum de son énergie libre. Cette hypothèse est connue sous le nom d'hypothèse thermodynamique [3], et prédire la structure naturelle d'une protéine soluble dans l'eau à partir de la séquence d'acides aminés reste toujours le problème non résolu du repliement de protéines.

Les techniques expérimentales, comme la cristallographie aux rayons X et la spectroscopy NMR peuvent déterminer avec précision la structure d'une protéine, mais ces techniques restent chères et coûteuses en temps. Ces données sont en accès libre à la Protein Data Bank (PDB) [6]. La détermination expérimentale de la structure d'une protéine peut demander jusqu'à plusieurs années en laboratoire. Nous avons alors beaucoup à gagner à simuler sur ordinateur les processus de repliement, prédisant, directement de sa séquence d'acides aminés, la structure d'une protéine.

Il a été observé que les protéines de différentes séquences partagent des formes naturelles très similaires et, par conséquent, le nombre des différentes conformations s'avère être étonnament limité [10]. Cette observation suggèrent que les protéines pourraient partager un comportement universel, indépendamment des détails chimiques au niveau primaire. Dans cette thèse, nous essayons de saisir les propriétés universelles des protéines à l'aide du concept d'invariance de jauge qui joue un rôle clé dans notre description des lois fondamentales de la physique. De cette façon, un modèle déterminé de manière quasi-unique a

été développé afin de décrire avec succès la phase d'effondrement de la protéine. Nous verrons par la suite que ce modèle s'appuie sur les solutions solitoniques de type *kink* qui sont des objets stables et peuvent être reliés aux motifs hélice-boucle-hélice dans les protéines.

Cette partie fournit une brève vue d'ensemble sur les protéines et les solitons en insistant particulièrement sur les propriétés les plus utiles pour les chapitres suivants. Nous présenterons dans un premier temps les bases concernant la structure des protéines, qui possède une hiérarchie bien organisée. Par la suite, nous passerons en revue les forces motrices du processus de repliement, ainsi que les hypothèses de mécanismes. Ensuite, nous introduirons la théorie des solitons, qui fournit la description du prototype du motif hélice-boucle-hélice. Enfin, nous donnerons un aperçu des articles dans la seconde partie.

Il existe de nombreuses revues et monographies sur la structures des protéines. Une des monographies est donné par Petsko & Ringe [48]. La plupart des informations de la partie concernant les protéines peuvent être retrouvées dans cet ouvrage.

Chapitre 2

Structure des protéines

Les protéines sont généralement divisées en trois classes : les protéines globulaires, les protéines membranaires et les protéines fibreuses. À partir de maintenant nous nous concentrerons sur les protéines globulaires, la classe la plus fréquente. Les protéines sont repliées, en général, d'après une hiérarchie structurelle bien organisée. La structure primaire est la séquence d'unités monomériques (aussi appelées résidus) le long de la chaîne polypeptidique. Au niveau secondaire, il existe deux types de structures localement régulières, soient les α-hélices et les β-feuillets. Ces éléments réguliers secondaires sont connectés par des boucles et assemblés pour former la structure tertiaire.

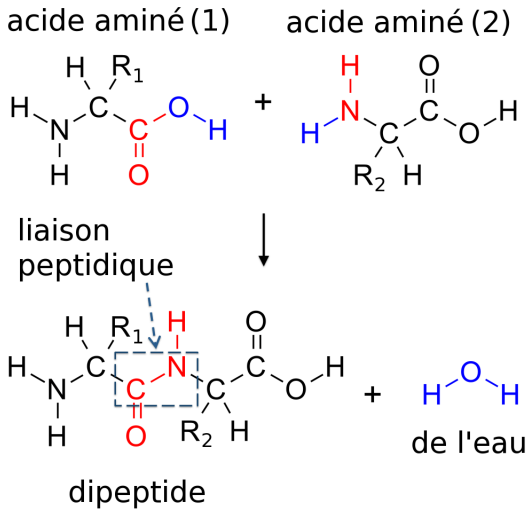

FIGURE 2.1 – Formation des liaisons peptidiques par processus de condensation de deux acides aminés. R_1, R_2 représentent les chaînes latérales.

2.1 Acides aminés et structure primaire

Les protéines appartiennent à la famille des hétéropolymères linéaires, composés d'un ensemble de vingt acides aminés. La séquence de ces résidus est encodée directement par le code génétique. Tous les acides aminés partagent une chaîne principale commune (Fig. 2.1), qui comprend un groupe amine (NH_2), un carbone alpha, C_α (le carbone auquel est attaché la chaîne latérale) et un groupe carboxyle (COOH). Les différents acides aminés se distinguent par leurs chaînes latérales qui sont responsables de leurs différentes propriétés chimiques. Les propriétés importantes comprennent la charge, l'hydrophilicité, l'hydrophobicité, la taille et les groupes fonctionnels.

Ces différentes propriétés vont fortement conditionner la formation de la structure protéique et les interactions protéine-protéine. On peut citer, à titre d'illustration, le cas des protéines solubles dans l'eau qui tendent à enfouir leurs résidus hydrophobiques à l'intérieur de leur configuration, tandis que les chaînes hydrophiles secondaires vont elles être préférentiellement exposées au solvant aqueux.La séquence d'acides aminés d'une protéine est connectée de manière covalente par des liaisons peptidiques. Ces dernières se forment entre un acide carboxylique COOH présent sur un acide aminé et un groupement amine NH2 d'un autre acide aminé, cette réaction libérant une molécule d'eau (Fig. 2.1). Du fait de cette liaison peptidique rigide, les atomes consécutifs C_α, C, N, C_α se situent dans un même plan et la distance entre les deux atomes C_α est fixe. Celle-ci est de 3.8 Angström dans la configuration trans, qui est la plus courante (2.8 Angström dans la configuration cis qui est elle beaucoup plus rare). Le dièdre impliquant C-N-C_α-C est appelée Φ alors que l'angle dièdre impliquant N-C_α-C-N est appelé Ψ. Ces deux angles (Ψ, Φ) ne sont pas contraints par la liaison peptidique et constituent donc les degrés de liberté de la chaîne protéique principale.

2.2 Structure secondaire

En se basant sur des considérations théoriques concernant la géométrie des protéines et les motifs des liaisons hydrogènes, Pauling a pu prédire la présence de structures secondaires régulières : les hélices α dextrogyres et les feuillets beta (Fig. 2.2). Il est important de noter la disposition régulière des liaisons hydrogènes à la fois dans les hélices α et dans les feuillets β. Au sein d'une hélice α, le groupe C=O de chaque résidu (n) forme une liaison hydrogène avec le groupe N-H du résidu (n+4), situé donc 4 résidus plus loin sur la séquence protéique. Dans le cas des feuillets β, deux brins ou plus peuvent être accolés spatialement et reliés par des liaisons hydrogène bien qu'étant très éloignés les uns des autres le long de la séquence de la protéine.

Il est possible de caractériser ces structures secondaires régulières à l'aide des angles dihedraux (Ψ, Φ) et de leur diagramme de distribution appelé aussi diagramme de Ramachandran (voir Fig. 2.3). Du fait des interférences stériques, il existe des régions interdite du diagramme de distribution des angles, aussi bien que des régions favorisées. Cette distribution est quasiment identique pour tous les résidus, à l'exception de la glycine qui possède plus de degrés de liberté du fait de son unique atome d'hydrogène, et de la proline qui elle possède moins de degrés de liberté à cause de sa structure cyclique.

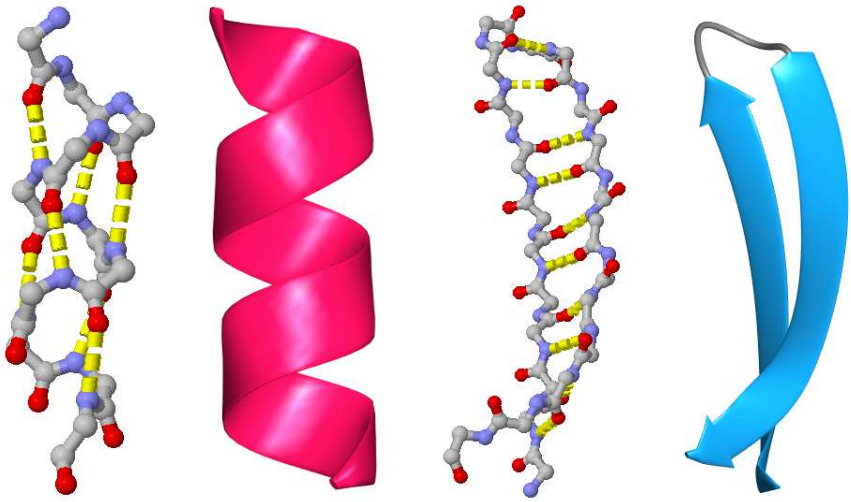

FIGURE 2.2 – (a)Diagramme schématique d'une hélice α. Les liaisons hydrogènes, représentées en jaune, sont disposées selon un motif régulier entre le résidu n et le résidu $n+4$. (b) Diagramme en ruban d'une hélice α. (c) Diagramme schématique d'un feuillet β. Les liaisons hydrogènes se forment entre deux brins. (d) Diagramme en ruban d'un feuillet β.

2.3 Structure tertiaire et domaines

Dans une protéine repliée, les hélices α et les feuillets β sont organisées dans un objet compact et presque solide, appelé la structure tertiaire de la protéine. Une séquence d'acides aminés dans la nature adopte une structure unique tertiaire, assurant la stabilité de sa fonction. Alors que les deux séquences similaires peuvent partager la ressemblance structurelle, il arrive aussi fréquemment que de nombreuses protéines ont des structures similaires, mais à faible similarité de séquence. Les protéines en tonneau TIM, qui appartiennent à structure tertiaire la plus courante, sont caractérisées par la structure un brin de feuillet β, suivi par une hélice α, répétée huit fois (voir Fig. 2.4). Parmi les membres de la famille tonneau TIM, il y a généralement un manque important d'homologie de séquence. En outre, ils ont aussi des fonctions très diverses.

Les grande protéines peuvent être spatialement décomposées en parties plus petites appelées domaines, qui sont des compacte régions globulaires, séparées par quelques acides aminés. D'autre part, deux ou plusieurs chaînes polypeptidiques peuvent s'associer pour former une structure complexe, appelée la structure quaternaire.

2.4 Représentation

A cause du grand nombre de degrés de liberté dans la structure d'une protéine, l'étude théorique devient extrêmement difficile pour la représentation atomique complete [56]. Par conséquent, un modèle réduit ou représentation à l'échelle moyenne est préférable, avec le compromis entre coût de calcul et précision. Grosso modo, les modèles à l'échelle moyenne

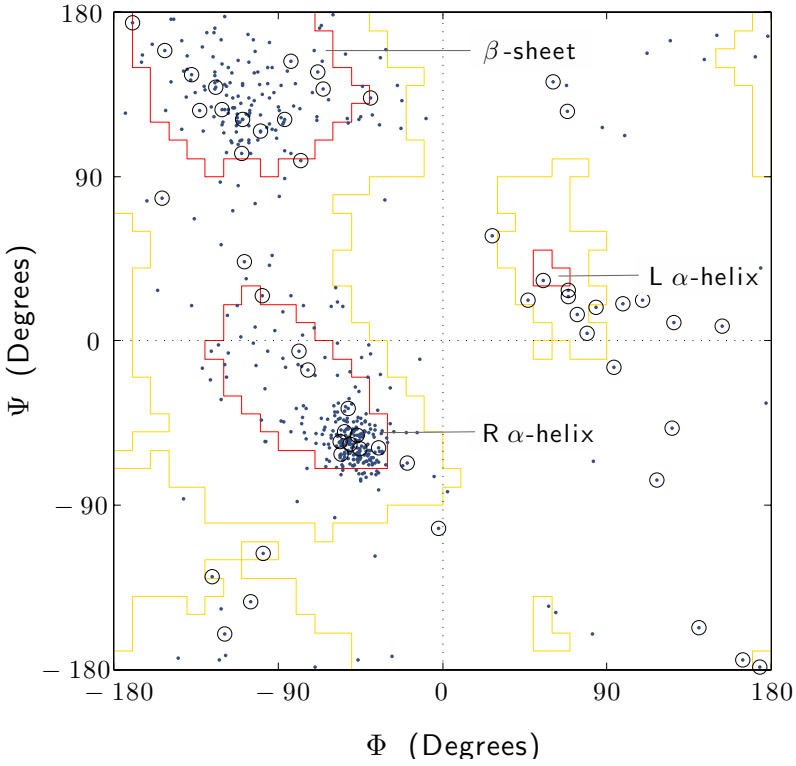

FIGURE 2.3 – Diagramme de Ramachandran pour 247 résidus de la protéine tonneau TIM (PDB 1tim). Les régions favorisées et permises sont délimitées respectivement par les lignes rouge et jaune. Les régions favorisées représentent les hélices α dextrogyres (R α-helix), feuillets β (β-sheet) et hélices α levogyres (L α-helix). Les régions permises, quant à elles, correspondent à des valeurs des angles psi/phi qui, bien que possibles, sont peu plausibles du fait de considérations énergétiques. Les limites de ces régions sont basées sur les calculs de Morris *et. al.* [44]. Les points encerclés correspondent au résidu de glycine.

FIGURE 2.4 – La structure tertiaire de 1TIM. Les hélices α sont de couleur rouge et les feuillets β bleu.

ont la même philosophie que l'approximation Born-Oppenheimer. La moyenne locale sur le mouvement rapide génère des représentations considérablement plus grossières que le détail atomique complet et permet la description du mouvement des protéines à des plus grandes échelles.

Traditionnellement, les études de la dynamique moléculaire (voir section 3.4) prennent en compte explicitement tous les atomes dans la protéine. Un autre choix courant est de n'inclure que les atomes lourds du squelette de la protéine (carbone et azote), parfois aussi incorporer le premier atome de la chaîne latérale (C_β) pour une meilleure modélisation des liaisons hydrogène. Comme une représentation même grossière, le C_α-squelette est en particulier utile et attrayant. Prendre l'avantage de distance constante entre atomes de carbone de 3.8 Å, la protéine entière est simplement représenté comme une chaîne d'atomes de C_α, avec le pseudo angle de lien ψ et le pseudo angle de torsion θ associé à chaque angle de C_α comme les degrés de liberté seulement. La structure tridimensionnelle d'une protéine de longueur N est complètement déterminée par l'ensemble des angles de liaison et les angles de torsion $\{(\psi_i, \theta_i), i = 1, \ldots, N\}$. Il a été démontré qu'il est possible de faire correspondre à un C_α-squelette une représentation de tous les atomes dans un système de protéines sans perte significative de l'information [27]. Analogue à la parcelle Ramachandran, la distribution de (ψ, θ) calculée à partir de PDBselect est aussi densément peuplée, indiquant les différents types de structure secondaire (Fig. 2.5).

Lorsque les protéines sont étudiées à l'ordinateur, il y a généralement deux choix, par

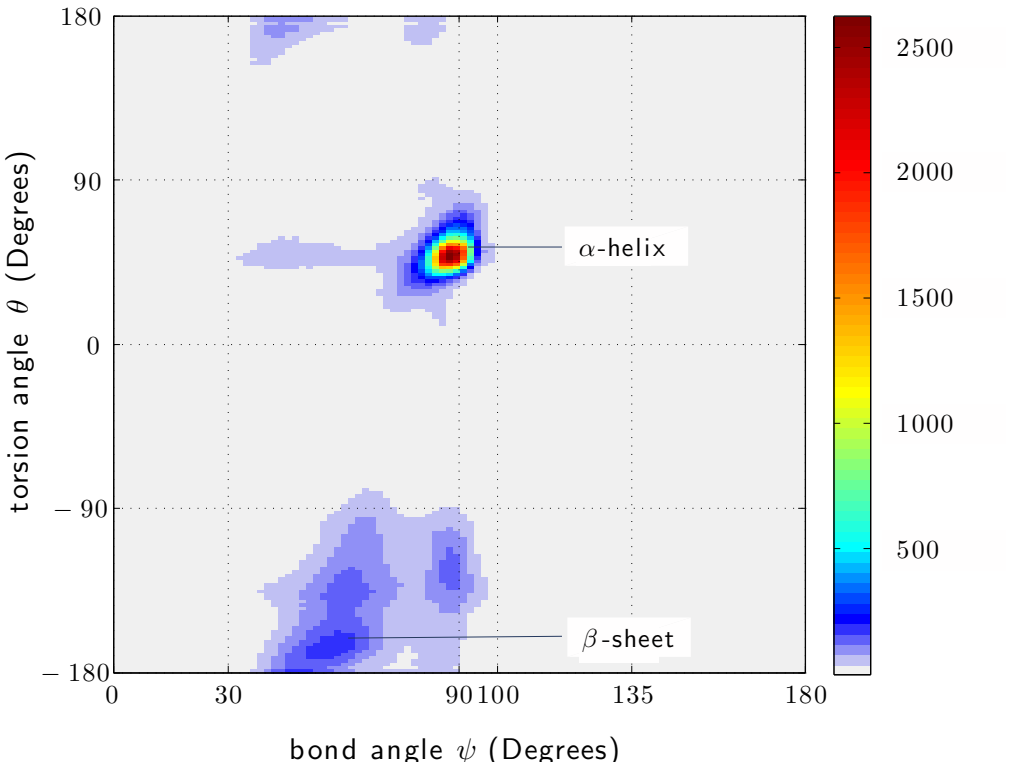

FIGURE 2.5 – La distribution des angles de liaison/torsion, calculée à partir de 2505 protéines non homologues du PDBselect (594 254 acides aminés). Les angles de liaison sont retenus pour la région de $[30°, 100°]$ à cause de l'encombrement stérique.

exemple les modèles sur réseau et hors-réseau. Dans les simulations sur réseau, les atomes sont limités à se déplacer sur un réseau prédéfini, ce qui discrétise l'espace tridimensionnel. Cela peut considérablement accélérer les simulations, mais au prix d'une perte considérable de précision. Comme la puissance de calcul des ordinateurs s'est améliorée ces dernières années, il devient plus populaire de choisir les simulations hors-réseau de protéines.

Chapitre 3

Le repliement des protéines

3.1 Paradoxe de Levinthal

En 1969, Levinthal [39, 63] a montré qu'étant donné trois états possibles pour chaque connection des acides aminés, une protéine de, disons, 101 acides aminés pourrait avoir $3^{100} \approx 5 \times 10^{47}$ configurations possibles. Même si la protéine pouvait assayer 10^{13}s^{-1} configurations par seconde, il lui faudrait $5{\times}10^{34}\text{s} \approx 10^{27}$ années pour choisir une quelconque parmi toutes. Cependant, beaucoup de petites protéines à un seul domaine ($\sim$110 acides aminés de longueur) se plient en moins d'une seconde ou même une milliseconde. Ce paradoxe suggère que les protéines ne réalisent pas de recherche aléatoire mais un cheminement dirigé. Tout modèle de repliement des protéines doit répondre à la résolution du paradoxe de Levinthal.

3.2 Les énergies

Outre les liaisons peptidiques dans les protéines, il y a parfois d'autres liaisons covalentes présentes sous la forme de ponts disulfure entre la chaîne latérale des résidus cystéine. Chacune de ces deux interactions covalentes a une énergie libre qui varie entre de 50kcal/mole et 150kcal/mole. Par contre, la valeur typique de variation d'énergie libre est

$$\Delta G = -15 \sim -5\text{kcal/mole} \tag{3.1}$$

de l'état déplié à l'état natif [47]. Apparemment, cette variation d'énergie libre ne vient pas de la contribution des liaisons covalentes, mais tombe dans la région des interactions non-covalentes (par exemple, une liaison hydrogène a une énergie libre de $1\sim$ 5kcal/mole). Ainsi, la structure complexe des protéines repliées est formée par l'interaction fragile entre les interactions non-covalentes. Toutes ces interactions non-covalentes polaires sont électrostatiques et l'effet est le même : des espèces polarisées positivement sont associées à celles polarisées négativement.

Une telle interaction électrostatique n'a été que récemment abordée dans un cadre de simulations de dynamique moléculaire quantique ab initio. La simulation de chaque petite peptide, cependant, exige un calcul très intensif. En conséquence, l'étude des protéines se

fait généralement dans le cadre en employant une fonctionnelle d'énergie semi-empirique. Il est également pratique de combiner ce calcul à celui plus traditionnel, des champs de forces par dynamique moléculaire ou des simulations Monte Carlo.

Les différents modes d'interactions non-covalentes sont généralement divisées en liaison hydrogène, interaction électrostatique à longue portée, et interaction de Van der Waals. En particulier, les liaisons hydrogène entre les protéines et les molécules d'eau, qui donnent lieu à effet hydrophobe, sont d'une grande importance pour la stabilité des protéines. Les résidus hydrophobes ou non-polaires préfèrent être enterrés à l'intérieur des protéines, afin de ne pas entrer en contact avec les molécules d'eau, tandis que les groupes hydrophiles, qui sont chargées ou polaires, resteront à la surface, en contact avec l'eau.

3.3 Mécanisme de repliement

Les deux simulations théoriques et les études expérimentales ont fourni des informations clés dans le mécanisme de repliement des protéines et dans la dynamique des différents états des protéines [17, 18]. Un modèle de mécanisme devrait expliquer comment une protéine recherche l'espace de configuration si vite, bien au-delà de la recherche aléatoire de Levinthal paradoxe.

Dans le modèle de diffusion-collision [5, 8, 35, 45], les éléments de structure secondaire locale formés (hélices α, feuilles β) dans le stade précoce se diffusent et se heurtent, se formant tard la configuration tertiaire. Le modèle d'effondrement hydrophobe [15, 49] soutient que les chaînes latérales hydrophobes fusionnent rapidement formant un noyau naissant de lointains contacts tertiaires. Sur ce noyau hydrophobe, les structures locales secondaires se propagent. Contrairement à ces deux modèles, le concept de nucléation [55] de recherche a l'idée différente qu'après l'achèvement du nombre minimal de contacts formant le noyau que l'on appelle le pliage, l'état natif sera atteint rapidement. Une preuve expérimentale pour le modèle de nucléation de recherche ressort des études qui mutent des résidus en boucles. Ces boucles mutantes sont plus critiques à l'étape limitante que ces résidus trouvés souvent dans une structure secondaire contigus, et donc, une boucle qui définit la topologie est plus limitante que la formation d'une structure simple secondaire des hélices ou des brins de feuille. Un autre mécanisme proposé est l'assemblage par étapes d'unités foldon [41].

Cependant, une telle proposition des mécanismes de pliage sont principalement des résumés d'expériences. Surtout, ils ne disent pas comment calculer l'itinéraire de pliage pour une protéine à partir de sa séquence d'acides aminés. En revanche, le mécanisme de compression et d'assemblage (ZA) prévoit un principe général qui pourrait prédire les routes de pliage et le taux de toute la séquence [25, 29, 31, 1, 57, 58, 59, 60, 61, 62]. En mécanisme de ZA, les fragments rechercher d'abord les peptidiques méta-structures stables locales. Mais peu d'entre eux sont suffisamment stables pour survivre pendant des échelles de temps plus long. En conséquence, soit ces structures augmentent (zip) dans des structures plus grandes et plus stables, ou soit assemblent avec d'autres structures. Il est alors facile de voir que le taux de pliage est limité par l'ordonnance de contact efficace (ECO) [22, 16]. Ce mécanisme peut donner la prévision du changement de l'itinéraire de pliage [60], ce qui est mesuré par la variation de la de distribution Φ-valeur [40].

3.4 Simulation

Comme la méthodologie et la puissance des ordinateurs est continuellement améliorée, des progrès ont été réalisés afin que les petites protéines de moins de 100 acides aminés puissent être prédites avec certitude [4, 32, 21, 7, 9, 19]. Mais il faudra encore du temps et des efforts pour simuler les protéines de plus de 150 acides aminés, qui jourent un rôle majeur en biologie. Il ya généralement deux classes de méthodologie, les méthodes basées sur les structures des protéines connues et l'approche *ab initio*. Pour la première, l'idée principale est basée sur l'observation que les protéines, qui sont très similaires, en ce qui concerne leurs séquences, le sont, également, en ce qui concerne leur facon de se réplier [34]. Ainsi, étant donné la protéine cible, ces méthodes de recherche de la structure emploient une base de données pour localiser des protéines à séquence similaire, et utilisent ces structures pour rapprocher la structure désirée. L'imprécision de ces méthodes, qui s'aggrave avec l'ecart entre l'identité des séquences de la protéine cible et celle du départ, provient des erreurs dans l'alignement des séquences initiales et de la sélection de modèle inapproprié [33]. Il est clair qu'avec une résolution expérimentale de plus en plus fine des structures, cette méthode a un potentiel encore plus grand dans l'avenir.

L'approche *ab initio*, au contraire, tente de résoudre le problème à partir de principes de base. Malgré le succès croissant des méthodes basées sur les structures des protéines connues , les méthodes *ab initio* resteront essentielles si nous voulons obtenir une image détaillée de la cinétique de repliement, ainsi que les relations entre la séquence protéique et celle de la structure. Il y a au moins deux défis distincts impliqués dans le problème de la prédiction *ab initio* de la structure. L'un est la conception d'une bonne fonction d'énergie, qui est généralement empirique et dépend fortement du choix de la représentation de la protéine . L'autre consiste à concevoir un échantillonnage efficace ou une stratégie de recherche de mouvement de protéines.

Selon l'hypothèse de thermodynamique postulé par Anfinsen, une protéines pliée arrive à un minimum de son énergie libre. Les techniques de minimisation principales dans l'étude théorique des molécules biologiques sont la dynamique moléculaire (DM) et la méthode de Monte Carlo (MC) [53]. la méthode de simulation MD est basée sur la deuxième loi de Newton,

$$\mathbf{F}_i = m_i \frac{d^2\mathbf{r}_i}{dt^2}, \mathbf{F}_i = -\frac{\partial}{\partial \mathbf{r}_i} E\left(\{\mathbf{r}_i\}\right), \tag{3.2}$$

où $\mathbf{F}_i$ est la force exercée sur l'atome i, m_i sa masse, $\mathbf{r}_i$ la coordonnée atome, et $E\left(\{\mathbf{r}_i\}\right)$ l'énergie libre de la protéine. En calculant les forces exercées sur l'atomes individueles (à partir de la fonction d'énergie), nous pouvons déterminer l'accélération de chaque atome dans le système. L'intégration des équations du mouvement par la suite mène à des résultats pour la trajectoire qui décrit la dynamique d'une protéine. A partir de cette trajectoire, nous pouvons alors obtenir les valeurs moyennes des propriétés, basé sur l'hypothèse ergodique, qui stipule que la moyenne sur la trajectoire est égale à la moyenne d'ensemble. La méthode est déterministe, par exemple, le positions initiales et les vitesses de chaque atome détermine l'état de la protéine à tout moment dans le futur ou le passé. Les simulations de DM peut être fastidieuse et coûteuse en ressources informatiques. Cependant, avec l'avance de la vitesse des ordinateurs, et de nouvelles méthodes telles que l'échange réplique MD, et des modèles réduits de protéines, il est maintenant possible de déterminer la structure

de l'état natif directement à partir de la séquence des petits protéines.

Plutôt que de modélisation de la dynamique d'un système, l'objectif d'une simulation de MC est de capturer des propriétés statistique (thermodynamique) d'un système par une recherche stochastique. Alors que les types de mouvements dans une simulation MD sont strictement dictés par les lois de Newton de la physique, il n'y a pas de telle restriction sur les mouvements dans une simulation MC. La seule exigence est que la simulation n'est pas biaisé, qui peut être assurée par l'application de bilan détaillé et ergodicité (voir Annexe A). En conséquence, la simulation MC améliore potentiellement la portée des simulations en termes de taille et le calendrier, et est donc largement utilisée pour des prédiction *ab initio* de la structure des protéines. La mise en œuvre populaire est fait avec le cadre de la chaîne de Markov MC, dans lequel l'équilibre génère la distribution de Boltzman du système de la protéine.

Chapitre 4

Solitons in proteins

There are multiple facets of soliton, one of which is here identified as the helix-loop-helix motif in folded proteins. In this section, we first review the general aspects of soliton theory, with emphasis on its application in protein theory. Then we give an overview of gauge theory in Frenet equations, for the purpose of introducing kink soliton solution. At last we present the prototype of kink soliton in the form of helix-loop-helix motif, with its biological interpretation discussed.

4.1 A glance at soliton

Just as the roles of harmonic oscillations and waves in linear physical models, the soliton-type localized excitations are fundamental to the problems in essentially nonlinear systems. John Scott-Russell first discovered the soliton phenomenon in 1844 [54], and further research led to understanding solitons as solutions to the Korteweg–de Vries equation, nonlinear Schrodinger equation, and Sine-Gordon equations [12, 42]. Its application has been widely found in physics, electronics, optics, technology and biology. Solitons of all these types cannot be obtained within the framework of a quasi-linear approach, that is, under the assumption that the linearized system is the first approximation. Instead, a soliton emerges when nonlinear interactions combine elementary constituents into a localized collective excitation that are stable against weak perturbations and behave like a quasi-particle with invariant shape and velocity. When two solitons collide, they merge into one and then separate into two with the same shape and velovity as before the collision. From the point of view of energy transfer, solitons are notably more efficient than linear waves. Especially, this manifests much explicitly in protein systems, in which we should take account of the anharmonicity due to hierarchy of interactions, anisotropy and flexibility of macromolecular chains.

For simplicity, take the one-dimensional system as the running example. If a one-dimensional system has at least two ground states, it is possible to construct a chain of strongly bound monomers whose left part is associated with one ground state, while the right part, with the other ground state, the transition region being localized. This is a clear

illustration of a kink-soliton similar, for example, to a solution of the double-well potential

$$\frac{d^2y(s)}{ds^2} = -\frac{dV(s)}{ds} = -\frac{d}{ds}\left[\frac{m^2}{2c^2}\left(y^2-c^2\right)^2\right], y(s) = c\tanh\left[m(s-s_0)\right]. \tag{4.1}$$

It describes a trajectory that interpolates between the two minima $y = \pm c$ of the potential $V(s)$.

The important thing is the physical interpretation of the state $y(s)$. In the protein research, the Davydov soliton describes a local structural change of the α-helix [13, 14]. As it has been shown previously, the hydrogen bonds that stabilize the structure of an α-helix have a regular pattern. The idea here is that the energy liberated in the hydrolysis of adenosine triphosphate (ATP) creates up to two quanta of amide-I, essentially a stretching vibration in the C=O bond. This vibration excitation propagates from one residue (residue n) to its neighbors ($n \pm 1$) by the dipole–dipole interaction. But it also interacts with the neighboring hydrogen bonds ($n \pm 4$), leading to a deformation of the regular pattern and a lower energy state. This new state, which traps the amide I oscillation energy and prevents its dispersion, is the Davydov soliton.

Our model gives another possibility of interpreting the soliton $y(s)$. Instead of the description of a local structural change, we associate $y(s)$ with the curvature of the C_α-backbone of the native protein. Solitons could then become indispensable in herding atomary level simulations to correctly capture the collective fluctuations that drive the folding process. To make it explicitly, we first review the gauge structure in the Frenet theory.

4.2 Gauge structure in the Frenet equations

The following description has been illustrated in [46] and Chapter 6. For completeness, we briefly list the key elements. In the continuum limit the protein backbone of length L can be approximated by a continuous one-dimensional string $\mathbf{r}(s)$ where the arc-length parameter $s \in [0, L]$. Since we are interested the shape of a protein, it is convenient to define a local Frenet frame that runs with the curve : the unit tangent vector $\mathbf{t}(s)$, the normal vector $\mathbf{n}(s)$ and the binormal vector $\mathbf{b}(s)$, which are given the relations

$$\mathbf{t} = \frac{d\mathbf{r}}{ds} \equiv \mathbf{r}_s, \mathbf{n} = \frac{\mathbf{t}_s}{|\mathbf{t}_s|}, \mathbf{b} = \mathbf{t}\times\mathbf{n}. \tag{4.2}$$

We then define a complex combination $\mathbf{e}_F^{\pm} = \frac{1}{2}(\mathbf{n} \pm i\mathbf{b})$. The derivatives of Frenet frame with respect to s are linked to the original frame by the Frenet equations

$$\frac{d\mathbf{t}}{ds} = \frac{1}{2}\kappa\left(\mathbf{e}_F^+ + \mathbf{e}_F^-\right), \frac{d\mathbf{e}_F^{\pm}}{ds} = -\kappa\mathbf{t} \mp i\tau\mathbf{e}_F^{\pm}, \tag{4.3}$$

with the curvature $\kappa(s)$ and the torsion $\tau(s)$ measuring the derivation of the string from straight line and plane curve, respectively. 'The concept of gauge invariance emerges from the following simple observation [46] : The vectors $\mathbf{n}$ and $\mathbf{b}$ span the normal plane of the string. But any physical property of the string must be independent of the choice of basis

on the normal plane' [11]. We could make a U(1) rotation with an angle $\eta(s)$ on the normal plane,

$$\mathbf{e}_F^{\pm} \to e^{\pm i\eta}\mathbf{e}_F^{\pm} \equiv \mathbf{e}_\eta^{\pm}, \tag{4.4}$$

and this sends

$$\kappa \to e^{i\eta}\kappa \equiv \kappa_\eta, \tau \to \tau - \partial_s\eta \equiv \tau_\eta. \tag{4.5}$$

So we can identify this relation as the gauge transformation structure of two dimensional Abelian Higgs multiplet : κ_η represents the complex scalar field while τ_η represents the spatial component of the U(1) gauge field. The gauge invariant principle then suggests a natural choice of the energy functional

$$E = \int_0^L ds \left\{ \left|\left(\partial_s - i\tau_\eta\right)\kappa_\eta\right|^2 + c\left(\left|\kappa_\eta\right|^2 - \mu^2\right)^2 \right\} + d\int_0^L ds\tau_\eta. \tag{4.6}$$

4.3 Soliton = helix-loop-helix motif

The discretization of the energy functional gives (with additional Proca mass term and a regulator term)

$$E = \sum_{i=1}^{N-1}\left(\kappa_{i+1} - \kappa_i\right)^2 + \sum_{i=1}^{N} c\left(\kappa_i^2 - m^2\right)^2 + \sum_{i=1}^{N}\left(b\kappa_i^2\tau_i^2 + d\tau_i + e\tau_i^2 + q\kappa_i^2\tau_i\right). \tag{4.7}$$

Unlike the conventional harmonic modeling with respect to the bond angle, this energy functional is essentially nonlinear. As a good approximation (verified by numerical computation), we firstly keep the first two sums and get the double-well ϕ^4 model that is known to support the topological kink soliton. In the continuum limit the kink soliton has the analytical form,

$$\kappa(s) = m\tanh\left(m\sqrt{c}\left(s - s_0\right)\right), \tag{4.8}$$

where s_0 is the central position of the soliton. On the other hand, variation with respect to the torsion results

$$\frac{\partial E}{\partial \tau_i} = 2b\kappa_i^2\tau_i + 2e\tau_i + d + q\kappa_i^2 = 0, \tag{4.9}$$

$$\tau_i = -\frac{1}{2}\frac{d + q\kappa_i^2}{e + b\kappa_i^2}. \tag{4.10}$$

These expressions of curvature and torsion, illustrated in Fig. 4.1, determines the shape of the string by solving the Frenet equations. The striking observation is that this kink soliton describes the profile of helix-loop-helix motif in protein. The two ending platforms of kink soliton corresponds to the helices while the transition region represents the loop connecting these two helices. On the other hand, there are totally six parameters in the model, capturing the six characteristics of helix-loop-helix, that is, the curvature and torsion of helices, the length of the loop (mainly tuned by parameter c), and the relative orientation between two helices (can be formulized by three Euler angles).

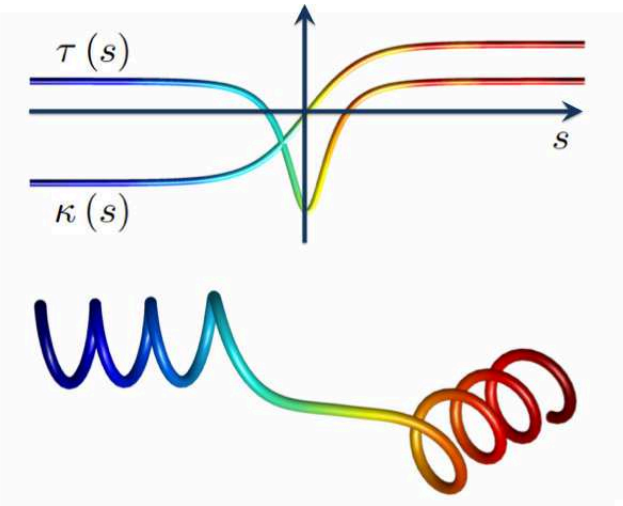

FIGURE 4.1 – Schematic relation between the kink soliton, the energy minimal solution, and the helix-loop-helix motif. The ground states correspond to helices while the transition region is mapped to the connecting loop. In practice, both soliton and protein structure are in their discrete version. See the publications for more details.

The discretization of the string, which is the result of the C_α-representation we take for practical proteins modeling [23], doesn't change the spirit of this beautiful correspondence between the kink soliton and helix-loop-helix. We also remark that the helix here is not only restricted to α-helix; β-sheets, another common secondary structure elements, can be regarded as the deformed helices because of their intrinsic twist. This prototype model further suggests the sequential connecting of multiple solitons will describe the longer proteins with more than two helices, similar with the idea of protein threading [33]. How to tune the parameters for the purpose of condensing the helices into compact tertiary structure is one of the main objects of this thesis (technical details are shown in appendix A).

This soliton model also suggests a partial solution of Levinthal paradox. The formation of secondary structure elements, equivalent to the soliton generation, proceeds in parallel and not in sequence as his thought experiment proposed. The stability of soliton ensures the integrality of secondary structure elements, which moves globally to collapse into compact conformation, driven by the hydrophobic effect. This global process speeds up the folding dramatically.

4.4 Biological interpretation of the energy functional

Though the energy functional is derived from principles of geometry and symmetry, it has a straightforward interpretation from the biological point of view. Firstly we point out that the discrete curvature and discrete torsion are synonymous with bond angle and torsion angle, respectively, as shown later in Chapter 6. The first sum in Eq. (4.7) accounts for the rigidness of bond angle, which can be justified by the statistics from PDB data. The

second sum, beyond the harmonic approximation, takes the nonlinearity into consideration. The third sum represents the coupling between bond angle and torsion angle, as well as the self-interaction of torsion angles.

One may wonder where is the amino acid sequence information in this energy functional. The answer is that sequence specific interaction has been encoded into the parameters. In other words, our model of energy functional can be taken as the effective description. Suppose we know an energy functional at the finer-grained level, $U(\mathbf{x}, \mathbf{X})$. For example, $\mathbf{x}$ represents variables of N, O, H atoms while $\mathbf{X}$ represents those of C_α atoms. The partial integral over $\mathbf{x}$ leaves us with an effective potential,

$$\beta U(\mathbf{X}) = -\log \int e^{-\beta U(\mathbf{x},\mathbf{X})} d\mathbf{x}, \tag{4.11}$$

in which the parameters take the contribution from the implicit interaction of $\mathbf{x}$ variables.

One similar example of partial integral of interaction is the translation of explicit solvent into implicit solvent. The huge number of solvent molecules can be averaged out to a single dielectric constant, ϵ_r. As a result, the interaction between two charges, q_i and q_j in the solution is effectively given by

$$U(q_i, q_j) \approx \frac{1}{\epsilon_r} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}, \tag{4.12}$$

with the factor of $1/\epsilon_r$ as the contribution of implicit solvent.

In our model, the six parameters sum over the contribution of amino acid interactions within a helix-loop-helix motif. From this point of view, the energy functional is working on the secondary structural level. This feature is absolutely not existing before in the protein research. While people sum over the contributions of interactions over each atom to fold the protein structure, our model represents the collective behavior of individual contribution.

The benefit of effective description is obvious. It greatly reduces the computation cost while largely keeps the essential characteristics of the energy functional. Just very recently, our model has been illustrated to successfully fold a protein (PDB 1yrf) into the right target structure within very short time [38].

Our approach may be applied to study special function-related structural motifs. In particular, the relation of soliton to the canonical structures in immunoglobulins could be of great interest [2]. On the other way, if we further establish the explicit relation between amino acid sequence and the parameters in our model, the present approach could be very interesting in protein structure modeling.

Finally we remark that though we identify the correspondence between soliton and helix-loop-helix motif, it remains unknown for the relation between the perturbation of soliton and the perturbation of protein structure, which would justify the stability of soliton behavior.

Chapitre 5

Overview of research papers

At the very beginning of my PhD study, my supervisor Antti J. Niemi, together with Ulf H. Danielsson and Martin Lundgren at Uppsala University, had begun to develop a gauge field theory of chirally folded homopolymers [11], which can catch the statistical properties of folded proteins in their collapse phase. In particular, they obtained homochirality by adding an apposite Chern-Simons term, commonly used in high energy physics to describe parity violation. At that time, the goal of our efforts was developing a more realistic description of the the protein structure in the native state.

The theory they obtained, taking the spirit of double-well ϕ^4 model, requires symmetric degenerate state for the curvature. For this purpose, the curvature has to be generalized to be signed, compared with the conventional case of positive value. Furthermore, the typical solition solution of ϕ^4 model, was expected to have potential application in proteins. These two aspects drove us to develop a consistent description of discrete Frenet equations, as well as the gauge symmetry. The success of this approach had also resolved the mysterious face of soliton, which turned out to closely connect with the helix-loop-helix motif in proteins. The main-chain folding of entire proteins can then be built from multiple solitons. With the help of Maxim Chernodub, the parameterization of our model was successfully resolved and these results led to my first publication.

While the collapse phase of proteins is characterized by the scaling law of gyration radius on the tertiary level, the regular structure of helix-loop-helix motif can be summarized globally beyond the chemical details, showing the existence of wide universality on the secondary level. Both observations parallel with the experimental fact that the number of protein structure is much limited compared with the number of protein amino acid sequence.

The picture looked very promising in principle. Yet the success of the whole story depended largely on the numerical simulation of folded proteins, which was not so straightforward. In my first publication, the so-called double optimization for the parameter fitting required random walk in both structure space and parameter space, resulting in the inefficient search. To improve the situation, we tried to find a *fast* way of approximating minimum solution given the parameters. For this purpose, Nora Molkenthin, together with Antti Niemi and me, introduced a novel generalization of the discrete nonlinear Schrodinger (GDNLS) equation, which is derived from the original energy functional. Since the poten-

tial in our model has two separate local minima, a known result of GDNLS in [28] ensured the existence of a dark soliton solution. And the algorithm they proposed had helped us greatly speed up the simulation. The result was summarized in my second publication.

The equivalence between helix-loop-helix and soliton had to be generalized in order to describe the more complex proteins, which usually have longer loops between α-helices and β-strands. In the language of soliton, this required two asymmetric ground states for the potential. One ground state was still associated with the α-helices while the other with β-strands, the deformed helices. The result thus provided more flexible description of protein structures and indeed performed very well on the test of the 153 amino acid myoglobin 1M6C, as well as of the helix-loop-strand segment 3DLK with longer loop. The collaboration with Andrei Krokhotin, Antti Niemi and Xubiao Peng brought my third publication.

Finally, it had been found necessary to revisit the transfer matrix formalism we had been using for discrete Frenet equation. Besides summarizing the basic concepts and formalism, Martin Lundgen, Antti Niemi and me had also worked out to include the information of C_β carbons in the discrete Frenet framing. All the results serve as the foundation of our approach and thus the corresponding publication is placed in the forefront of second part, though it was born lastly.

A short description of each chapter follows here. Detailed introductions are included in the individual papers, and a summary included in the chapter of *Conclusion*. Relating techniques are shown in details in appendices.

Chapter 6 : Discrete Frenet frame, inflection point solitons, and curve visualization with applications to folded proteins

This paper summarize the fundamental formalisms of Frenet equation and gauge symmetry in our study of protein folding. We utilize the transfer matrix formalism to consecutively map the discrete Frenet frame from one vertex of discrete curve to its neighbor. This intrinsically discrete approach enables us to conveniently describe curves, the backbone of folded proteins for example, for which the continuum limit has a nontrivial Hausdorff dimension. It is also found that in the case of differentiable curves the continuum limit of our discrete equation reproduces the classical Frenet equation. Meanwhile, the inflection points, where the curvature vanishes and classical Frenet equation becomes invalid, are shown to be similar with topologically stable solitons. As an application, we consider the relative orientation of the C_β carbons with respect to the ensuring backbone C_α carbons. Statistics shows that both the relative orientation of the C_β carbons and the C_α-backbone geometry have almost identical indicators of the secondary structure preference.

Chapter 7 : Topological solitons and folded proteins

With the aid of local $\mathbb{Z}_2$ gauge symmetry, we uncover the soliton configurations in native protein structure. The α-helices (β-sheets can be treated as deformed helices) are viewed as ground states which are connected by the loops as solitons. We also present an energy functional that supports soliton solutions. We apply the energy functional to model

several proteins, by two ways of parameter fitting. The first way is to approximate the exact solution of the equations of motion by a continuum kink soliton solution. The second way is to optimize an objective functional that guides the parameter flow according to the difference between minimum structure and native structure. Both methods describe folded proteins within experimental accuracy.

Chapter 8 : Discrete nonlinear Schrodinger equation and polygonal solitons with applications to collapsed proteins

The energy functional in the previous work is further shown to be related with a novel generalization of the discrete nonlinear Schrodinger equation. It supports a dark soliton solution which can be fast computed by a fixed point method [28], compared to the Monte Carlo search in the last paper. The fitting with respect to protein structure also generates better accuracy.

Chapter 9 : Towards quantitative classification of folded proteins in terms of elementary functions

We generalized the soliton solution into an asymmetric form, to endow it with more flexibility on modeling protein structure. Using the myoglobin 1M6C as an example, we demonstrate that the native fold of a relatively long protein emerges from a combination of generalized soliton solutions with a high atomary-level accuracy. We also verify that our approach can be extended to longer loops and loops connecting α-helices and β-strands.

Deuxième partie

Research Papers

Chapitre 6

Discrete Frenet frame, inflection point solitons, and curve visualization with applications to folded proteins

Shuangwei Hu, Martin Lundgren, and Antti J. Niemi

Physical Review E 83, 061908 (2011)

Discrete Frenet frame, inflection point solitons, and curve visualization with applications to folded proteins

Shuangwei Hu,[1,2] Martin Lundgren,[1] and Antti J. Niemi[1,2]

[1]*Department of Physics and Astronomy, Uppsala University, P. O. Box 803, S-75108 Uppsala, Sweden*

[2]*Laboratoire de Mathematiques et Physique Theorique CNRS UMR 6083, Fédération Denis Poisson, Université de Tours, Parc de Grandmont, F-37200 Tours, France*

(Received 28 February 2011; published 15 June 2011)

We develop a transfer matrix formalism to visualize the framing of discrete piecewise linear curves in three-dimensional space. Our approach is based on the concept of an intrinsically discrete curve. This enables us to more effectively describe curves that in the limit where the length of line segments vanishes approach fractal structures in lieu of continuous curves. We verify that in the case of differentiable curves the continuum limit of our discrete equation reproduces the generalized Frenet equation. In particular, we draw attention to the conceptual similarity between inflection points where the curvature vanishes and topologically stable solitons. As an application we consider folded proteins, their Hausdorff dimension is known to be fractal. We explain how to employ the orientation of C_β carbons of amino acids along a protein backbone to introduce a preferred framing along the backbone. By analyzing the experimentally resolved fold geometries in the Protein Data Bank we observe that this C_β framing relates intimately to the discrete Frenet framing. We also explain how inflection points (a.k.a. soliton centers) can be located in the loops and clarify their distinctive rôle in determining the loop structure of folded proteins.

DOI: 10.1103/PhysRevE.83.061908 PACS number(s): 87.15.A−, 87.15.bd

I. INTRODUCTION

The visualization of a three-dimensional discrete framed curve is an important and widely studied topic in computer graphics, from the association of ribbons and tubes to the determination of camera gaze directions along trajectories. Potential applications range from aircraft and robot kinematics to stereo reconstruction and virtual reality [1,2].

We are interested in addressing the problem of characterizing the physical laws that govern protein folding. For this we develop a technique for framing a general discrete and piecewise linear curve. Our goal is to combine the geometric problem of framing with an appropriate physical principle for frame determination. Ultimately we hope to have an approach, where instead of purely geometric considerations the frames along a curve are determined directly from the properties of an underlying physical system. As a consequence we expect that our formalism and our results will find wide applicability well beyond the protein folding problem, where the present formalism has already found several applications [3–6].

The classical theory of continuous curves in three-dimensional space employs the Frenet equation [1,2] to determine a moving coordinate frame along a sufficiently differentiable space curve. However, if the curve has inflection points and/or straight segments or if it fails to be at least 3 times continuously differentiable, the Frenet frame becomes either discontinuous or may not even exist. In such cases there can be good reasons to consider the option to introduce an alternative framing such as Bishop's parallel transport frame [7], a geodetic reference frame, or some possibly hybrid variants [1,2].

In this article we derive a discrete version of the Frenet equation that introduces a framing along an intrinsically discrete and piecewise linear curve in $\mathbb{R}^3$. We develop the general formalism for the visualization of such a curve without any underlying assumption that it approaches a continuous space curve in the limit where the maximum length of its line segments goes to zero. The continuum limit may as well be a fractal, with a nontrivial Hausdorff dimension. Thus, unlike in several approaches that we are aware of, our starting point is not in a discretization of the continuum Frenet equation. Instead our approach is intrinsically discrete, and it is based on the transfer matrix formalism that is widely used for example in lattice field theories [8]. Indeed, we find it useful to adapt some notions of lattice gauge theories [8]. For us this provides a valuable conceptual point of view. Moreover, the transfer matrix formalism intrinsically incorporates self-similarity and thus the very concept of line segment length has no role in our derivations. Consequently, we can effortlessly consider curves that have fractal continuum limits, while at the same time ensuring that, if the continuum limit exists as a class $\mathcal{C}^3$ space curve, we recover the standard Frenet framing together with its generalized versions.

As an application we consider folded proteins, for which the continuum limit is known to be a fractal with Hausdorff dimension that is very close to three [3]. The locations of the central C_α carbon atoms along the protein determines a discrete piecewise linear curve; this is the protein backbone. We introduce a framing to the backbone by employing the C_β carbon atoms of the side chain amino acids that are covalently bonded to the C_α carbons that define the backbone. The frame at the location of a given C_α carbon is determined by the directional vector that connects it with the ensuing C_β carbon, together with the directional vector that connects it to the next C_α carbon along the backbone. By inspecting the framing of all protein structures in the Protein Data Bank (PDB) [9] we find that such a C_β framing relates intimately to the discrete Frenet framing of the backbone. In particular, we conclude that for a folded protein the concept of an inflection point acquires an intrinsic biological interpretation; it coincides with the location

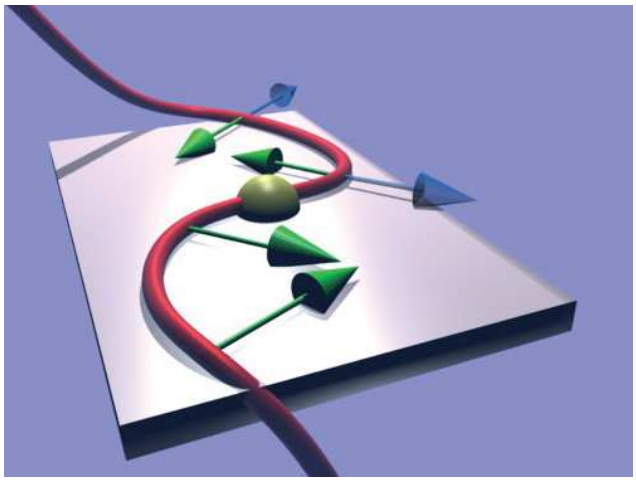

FIG. 1. (Color online) A curve with inflection point (ball). At each point the direction of the (Frenet frame) normal vectors is toward the center of an oscullating circle. There is a discontinuity in the direction of the normal vectors when we traverse the inflection point. At this point the radius of the oscullating circle diverges and the normal vector **n** becomes abruptly reflected in the oscullating plane from one side to the other side of the curve. The direction of the ensuing vector is opposite to the (reflected) normal vector **n** (see also Fig. 6).

of the center of the loop in a folded protein. Indeed, these inflection points drive the protein loop geometry, an isolated inflection point is topologically stable and it cannot be removed by any local continuous deformation of the curve. We remark that this kind of topological stability is inherent to solitons such as the kink-soliton and propose that the concept of solitons is a profitable one to understand the folding of proteins.

This connection between inflection points and topological solitons such as the kink can be understood as follows: At an isolated inflection point of a continuous curve, the curvature that is a frame-*independent* geometric characteristic of the curve vanishes. At such a point the Frenet frame can become discontinuous (see Fig. 1).

Consequently, a single nondegenerate inflection point cannot be removed by any local continuous deformation of the curve. An isolated nondegenerate inflection point can be only locally and continuously removed in the presence of another inflection point by deforming the curve so the inflection points annihilate each other in a saddle-node bifurcation. In particular, a sole nondegenerate inflection point can be removed only by translating it away through an end point of the curve that involves a *global* deformation of the curve. This kind of stability enjoyed by an isolated inflection point under local deformations of the curve is the hallmark of a topological soliton. Indeed, let us recall the topological kink-soliton in a quartic double-well potential [10]

$$\ddot{y} = -\frac{d}{ds}V(s) = -\frac{d}{ds}\left[\frac{m^2}{2c^2}(y^2-c^2)^2\right] = -\frac{2m^2}{c^2}y(y^2-c^2)$$
$$y(s) = c\tanh[m(s-s_0)]. \qquad (1)$$

It describes a trajectory that interpolates between the two minima $y = \pm c$ of the potential $V(s)$; see Fig. 2.

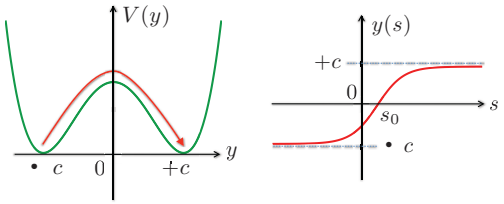

FIG. 2. (Color online) The kink-soliton (right) interpolates between the two ground states at $\phi = \pm c$ of the potential (left) as $s \to \pm\infty$. It is topologically stable and cannot be removed by any finite energy deformation.

The center of the soliton is at the point $s = s_0$ where $y(s)$ vanishes. The influence of this center point to the global topology of the trajectory cannot be removed by any kind of continuous local deformation $y(s) \to y(s) + \delta y(s)$, as the resulting curve continues to retain its characteristic global property that $y \to \pm c$ as $s \to \pm\infty$. Thus the deformed $y(s)$ necessarily vanishes at least at one point. The goal of the present paper is to explain how this signature behavior of a topological soliton can be detected and described in the case of discrete piecewise linear curves and in particular those curves that relate to the framing of folded proteins.

II. THE GENERALIZED FRENET FRAME, INFLECTION POINTS, AND SOLITONS

A. The generalized Frenet frame

We start by describing the continuum Frenet equation and its generalizations. Let $\mathbf{x}(s)$ be a space curve in $\mathbb{R}^3$. Its unit tangent vector

$$\mathbf{t} = \frac{1}{||\dot{\mathbf{x}}||}\dot{\mathbf{x}} \equiv \frac{1}{||\dot{\mathbf{x}}||}\frac{d\mathbf{x}(s)}{ds}$$

(we assume that $||\dot{\mathbf{x}}|| \neq 0$) is subject to the Frenet equation [1,2]

$$\frac{d}{ds}\begin{pmatrix}\mathbf{n}\\ \mathbf{b}\\ \mathbf{t}\end{pmatrix} = ||\dot{\mathbf{x}}||\begin{pmatrix}0 & \tau & -\kappa\\ -\tau & 0 & 0\\ \kappa & 0 & 0\end{pmatrix}\begin{pmatrix}\mathbf{n}\\ \mathbf{b}\\ \mathbf{t}\end{pmatrix}, \qquad (2)$$

where

$$\mathbf{b} = \frac{\dot{\mathbf{x}}\times\ddot{\mathbf{x}}}{||\dot{\mathbf{x}}\times\ddot{\mathbf{x}}||}$$

is the unit binormal vector and

$$\mathbf{n} = \mathbf{b}\times\mathbf{t}$$

is the unit normal vector of the curve, and

$$\kappa(s) = \frac{||\dot{\mathbf{x}}\times\ddot{\mathbf{x}}||}{||\dot{\mathbf{x}}||^3}$$

is the *frame independent* curvature of $\mathbf{x}(s)$ and

$$\tau(s) = \frac{(\dot{\mathbf{x}}\times\ddot{\mathbf{x}})\cdot\dddot{\mathbf{x}}}{||\dot{\mathbf{x}}\times\ddot{\mathbf{x}}||^2}$$

is the torsion. The three vectors $(\mathbf{n},\mathbf{b},\mathbf{t})$ form the right-handed orthonormal Frenet frame at each point of the curve.

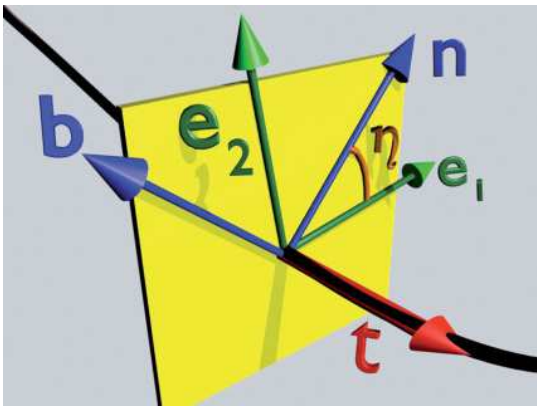

FIG. 3. (Color online) The Frenet frame $(\mathbf{n},\mathbf{b})$ and a generic orthogonal frame $(\mathbf{e}_1,\mathbf{e}_2)$ on the normal plane of $\mathbf{t}$, the tangent vector of the curve.

In the following we shall assume with no loss of generality, that $s \in [0,L]$ measures the proper length along a curve with total length L in $\mathbb{R}^3$ so

$$||\dot{\mathbf{x}}|| = 1. \tag{3}$$

Consider a curve with an isolated nondegenerate inflection point (or, more generally, a straight segment) such as the one depicted in Fig. 1. At the inflection point $s = s_0$ the Frenet frame cannot be introduced since $\kappa(s_0)$ vanishes; in the proper length gauge

$$\kappa(s_0) = ||\ddot{\mathbf{x}}(s_0)|| = 0.$$

Conventionally, see, e.g., Ref. [11], in the presence of inflection points, the Frenet equation (2) is usually introduced only piecewise between the inflection points for those values of s for which $\kappa(s)$ is nonvanishing. But there are also alternative approaches that allow for a continuous passage of the frame through the inflection point (more generally straight segments). For this we view the Frenet frame as an example of a general frame, obtained by starting from the observation that while the tangent vector $\mathbf{t}(s)$ for a given curve is unique, instead of $\{\mathbf{n}(s),\mathbf{b}(s)\}$ we may choose an arbitrary orthogonal basis $\{\mathbf{e}_1(s),\mathbf{e}_2(s)\}$ for the normal planes of the curve that are perpendicular to $\mathbf{t}(s)$, without deforming the curve. This general frame is related to the Frenet frame by a local SO(2) frame rotation around the frame-independent tangent vector $\mathbf{t}(s)$ (see Fig. 3),

$$\begin{pmatrix}\mathbf{n}\\ \mathbf{b}\end{pmatrix} \to \begin{pmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{pmatrix} = \begin{bmatrix}\cos\eta(s) & -\sin\eta(s)\\ \sin\eta(s) & \cos\eta(s)\end{bmatrix}\begin{pmatrix}\mathbf{n}\\ \mathbf{b}\end{pmatrix}. \tag{4}$$

The ensuing rotated version of the Frenet equation is

$$\frac{d}{ds}\begin{pmatrix}\mathbf{e}_1\\ \mathbf{e}_2\\ \mathbf{t}\end{pmatrix} = \begin{bmatrix}0 & (\tau-\dot{\eta}) & -\kappa\cos\eta\\ -(\tau-\dot{\eta}) & 0 & -\kappa\sin\eta\\ \kappa\cos\eta & \kappa\sin\eta & 0\end{bmatrix}\begin{pmatrix}\mathbf{e}_1\\ \mathbf{e}_2\\ \mathbf{t}\end{pmatrix}. \tag{5}$$

If we recall the adjoint basis of SO(3) Lie algebra

$$T^1 = \begin{pmatrix}0&0&0\\0&0&-1\\0&1&0\end{pmatrix}\quad T^2 = \begin{pmatrix}0&0&1\\0&0&0\\-1&0&0\end{pmatrix}\quad T^3 = \begin{pmatrix}0&-1&0\\1&0&0\\0&0&0\end{pmatrix},$$

where

$$[T^a,T^b] = \epsilon^{abc}T^c,$$

we find that on τ and κ the SO(2) transformation acts as follows,

$$\tau \to \tau - \dot{\eta} \tag{6}$$

$$\kappa T^2 \to \kappa(T^2\cos\eta - T^1\sin\eta) \equiv e^{\eta T^3}(\kappa T^2)\,e^{-\eta T^3}. \tag{7}$$

If instead of $\eta \equiv 0$ that specifies the Frenet frame (Frenet gauge) we select $\eta(s)$ so

$$\eta(s) = \int_0^s \tau(s')ds',$$

we arrive at Bishop's parallel transport frame [1,2,7] that can be defined continuously and unambiguously through inflection points. We note that (6) and (7) can be interpreted in terms of a SO(2) gauge multiplet [12]: The change (6) in $\tau(s)$ is identical to the SO(2) $\simeq U(1)$ gauge transformation of a one-dimensional gauge vector while $\kappa(s)$ transforms like a component of a SO(2) scalar doublet. This leads us to a gauge-invariant quantity, the complex valued Hashimoto variable [13]

$$\xi(s) = \kappa(s)\exp\left(i\int_0^s \tau\,ds'\right). \tag{8}$$

When we combine (6) with a SO(2) $\subset$ SO(3) rotation (7) by $\eta(s)$ around the T^3 direction of the SO(3) Lie algebra, the effect on (8) can be summarized as follows:

$$\xi(s) \to \left[\kappa(s)e^{-i\eta(s)}\right]\left\{\exp\left[i\int_0^s \tau\,ds' + i\eta(s)\right]\right\}e^{i\eta(0)} \tag{9}$$

and thus the Hasimoto variable $\xi(s)$ is manifestly independent of $\eta(s)$. (Note, however, that the $\eta(0)$ dependence remains as an overall global phase ambiguity that is inherent to (9); the local gauge invariance becomes eliminated but a global one remains.) In fact, the Hasimoto variable simply combines the two real components of the SO(2) scalar doublet into a single complex valued variable, with modulus that equals the frame independent (a.k.a., gauge-invariant geometric curvature of the curve. In particular the Frenet frame is like the widely used "unitary gauge" in the Abelian Higgs model [12].

We find this language of gauge transformations in connection of frame rotations introduced in Ref. [12] to be intuitively appealing and beneficial, and we shall use it frequently in the sequel.

B. Inflection points

We proceed to consider a continuous curve with n inflection points at $s = s_i$,

$$s_0 = 0 < \cdots < s_i < s_{i+1} < \cdots < L = s_{n+1}.$$

For simplicity, we assume that the inflection points are isolated and nondegenerate zeros of the curvature

$$\kappa(s_i) = 0.$$

A generalization to more involved inflection points is straightforward. We take the curve to be of class $\mathcal{C}^3$. This ensures that at each segment (s_i,s_{i+1}) the curvature is of class $\mathcal{C}^1$.

Furthermore, since the inflection points are nondegenerate, as we approach an inflection point the left and right derivatives of the curvature are nonvanishing and in the limit when $s \to s_i$ they become equal in magnitude but have an opposite sign,

$$\frac{d\kappa(s)}{ds}_{|s_i^+} = -\frac{d\kappa(s)}{ds}_{|s_i^-} \neq 0.$$

This jump in the derivative of the curvature is the signature of an inflection point in the Frenet frame. But even though the curvature $\kappa(s)$ fails to be continuously differentiable, the signed curvature

$$\tilde{\kappa}(s) = \sum_{i=0}^{n} (-1)^i \kappa(s)\theta(s - s_i)\theta(s_{i+1} - s) \tag{10}$$

with $\theta(s)$ the unit step function

$$\theta(s) = \begin{cases} 1 & s > 0 \\ 0 & s < 0 \end{cases}$$

is now continuously differentiable for all $s \in [0, L]$ and, in particular,

$$\frac{d\tilde{\kappa}}{ds}_{|s_i} \neq 0.$$

The original Frenet curvature $\kappa(s)$ and the signed curvature $\tilde{\kappa}(s)$ are related by a gauge transformation (7) of the Frenet frame, with $\eta(s)$ given by the following gauge transformation (6) of the Frenet torsion

$$\begin{aligned} \tau(s) \to \tau(s) - \dot{\eta}(s) &= \tau(s) - \pi \frac{d}{ds} \sum_{i=1}^{n-1} \theta(s - s_i) \\ &= \tau(s) - \pi \sum_{i=1}^{n-1} \delta(s - s_i). \end{aligned} \tag{11}$$

This can be immediately verified by comparing the form of (10) with that of the Hashimoto variable (8) and (9). We may call this gauge transformed version of the Frenet frame the $\mathbb{Z}_2$ Frenet frame, its discrete version will become important to us when we consider applications to folded proteins.

C. Solitons

For a concrete example of an inflection pint, we take the plane curve in Fig. 1. For this curve, in the vicinity of the inflection point the Frenet curvature has clearly a qualitative form that may be described by the absolute value of the kink-soliton profile (1),

$$\kappa(s) \sim \kappa_0 \left|\tanh[m(s - s_0)]\right|.$$

Obviously the derivative of this curvature is discontinuous with a finite jump at the inflection point I where $s = s_0$. This discontinuity reflects itself in the abrupt change in the direction of the (green) normal vector $\mathbf{n}$, as depicted in Fig. 1. The ensuing *signed* curvature (10) is qualitatively described by the kink-soliton (1)

$$\tilde{\kappa}(s) \sim \kappa_0 \tanh[m(s - s_0)] \tag{12}$$

and it is manifestly continuously differentiable, including the point $s = s_0$. Now the direction of the corresponding normal vector is also continuous through the inflection point. This is because the change in its direction becomes compensated by the change in the sign of the signed curvature when we cross the inflection point; see the blue vectors in Fig. 1 and Fig. 6. The example clearly exhibits the intimate relation between the concepts of inflection point and topological soliton.

III. THE DISCRETE FRENET EQUATION

A. The discrete Frenet frame

In the sequel we are primarily interested in an open and oriented, piecewise linear discrete curve that we describe by a three-vector $\mathbf{r}(s) \in \mathbb{R}^3$. The parameter $s \in [0, L]$ measures the arc length and L is the total length of the curve. The curve is determined by its vertices C_i that are located at the positions $\mathbf{r}_i = (\mathbf{r}_0, \ldots, \mathbf{r}_n)$ with $\mathbf{r}(s_i) = \mathbf{r}_i$. The end points of the curve are at $\mathbf{r}(0) = \mathbf{r}_0$ and $\mathbf{r}(L) = \mathbf{r}_n$. The nearest-neighbor vertices C_i and C_{i+1} are connected by the line segments

$$\mathbf{r}(s) = \frac{s - s_i}{s_{i+1} - s_i} \mathbf{r}_{i+1} - \frac{s - s_{i+1}}{s_{i+1} - s_i} \mathbf{r}_i$$

where $s_i < s < s_{i+1}$. We utilize the Galilean invariance to translate the base of the curve to the origin in $\mathbb{R}^3$ so

$$\mathbf{r}_0 = 0.$$

The remaining global rotational orientation of the curve can then be fully determined by the choice of $\mathbf{r}_1$ and $\mathbf{r}_2$.

For each pair of nearest-neighbor vertices $\mathbf{r}_{i+1}$ and $\mathbf{r}_i$ along the curve we introduce the unit tangent vector

$$\mathbf{t}_i = \frac{\mathbf{r}_{i+1} - \mathbf{r}_i}{|\mathbf{r}_{i+1} - \mathbf{r}_i|}. \tag{13}$$

If all tangent vectors are known, the position of the kth vertex is given by

$$\mathbf{r}_k = \sum_{i=0}^{k-1} |\mathbf{r}_{i+1} - \mathbf{r}_i| \cdot \mathbf{t}_i. \tag{14}$$

We now introduce the discrete Frenet frame (DF frame) at the vertex C_i at $\mathbf{r}_i$. This can be done whenever the three vertices at $\mathbf{r}_{i+1}$, $\mathbf{r}_i$, and $\mathbf{r}_{i-1}$ are not located on a common line so $\mathbf{t}_i$ and $\mathbf{t}_{i-1}$ are not parallel. This enables us to determine the unit binormal vector

$$\mathbf{b}_i = \frac{\mathbf{t}_{i-1} \times \mathbf{t}_i}{|\mathbf{t}_{i-1} \times \mathbf{t}_i|} \quad (i = 1, \ldots, n-1) \tag{15}$$

and the unit normal vector

$$\mathbf{n}_i = \mathbf{b}_i \times \mathbf{t}_i. \tag{16}$$

The orthogonal triplet $(\mathbf{n}_i, \mathbf{b}_i, \mathbf{t}_i)$ constitutes the discrete Frenet frame (DF frame) for the curve at the position of the vertex $\mathbf{r}_i$ for each $i = (1, \ldots, n-1)$; see Fig. 4.

B. The transfer matrix

We now proceed to derive a discretized version of the Frenet equation (DF equation) that relates the discrete Frenet frame at vertex C_i to the discrete Frenet frame at vertex C_{i+1} and allows for the construction of the curve in terms of the appropriate discrete versions of the curvature $\kappa(s)$ and torsion $\tau(s)$.

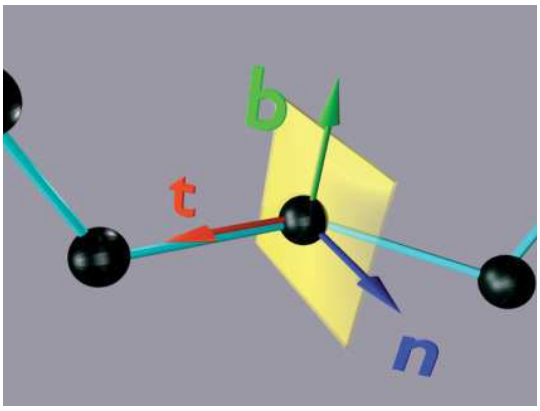

FIG. 4. (Color online) A discrete piecewise linear curve is defined by its vertices C_i and at each vertex there is an orthonormal discrete Frenet frame $(\mathbf{t}_i,\mathbf{n}_i,\mathbf{b}_i)$, provided $\mathbf{t}_{i-1}$ and $\mathbf{t}_i$ are not parallel.

From general considerations [8] we conclude that the DF equation should involve a transfer matrix $\mathcal{R}_{i+1,i}$ that maps the DF frame at the vertex i to the DF frame at the vertex $i+1$,

$$\begin{pmatrix}\mathbf{n}_{i+1}\\ \mathbf{b}_{i+1}\\ \mathbf{t}_{i+1}\end{pmatrix} = \mathcal{R}_{i+1,i}\begin{pmatrix}\mathbf{n}_i\\ \mathbf{b}_i\\ \mathbf{t}_i\end{pmatrix}. \tag{17}$$

The construction of this transfer matrix then amounts to a solution of the DF equation:

$$\begin{pmatrix}\mathbf{n}_n\\ \mathbf{b}_n\\ \mathbf{t}_n\end{pmatrix} = \mathcal{R}_{n,n-1}\cdot\mathcal{R}_{n-1,n-2}\cdot\ldots\cdot\mathcal{R}_{2,1}\begin{pmatrix}\mathbf{n}_1\\ \mathbf{b}_1\\ \mathbf{t}_1\end{pmatrix}$$

so once the transfer matrix is known for all $i=1,\ldots,n-1$, we can use (17) to construct all the Frenet frames for $i=2,\ldots,n$ and the entire curve $\mathbf{r}(s)$ using (14) together with the fact that the curve is linear in the intervals $s_{i-1}<s<s_i$. We recall that for the initial conditions we need to specify $\mathbf{r}_0$ that we have already chosen to coincide with the origin $\mathbf{r}_0=0$ and $\mathbf{r}_1$ and $\mathbf{r}_2$ that remove the degeneracy under global SO(3) rotations of the curve in $\mathbb{R}^3$.

The transfer matrix $\mathcal{R}_{i+1,i}$ is an element of the adjoint representation of SO(3), and thus we can parametrize it in terms of Euler angles. We choose the (zxz) angles

$$\begin{aligned}
\mathcal{R}^{11}_{i+1,i} &= -\sin\psi\sin\phi+\cos\theta\cos\psi\cos\phi\,|_{i+1,i}\\
\mathcal{R}^{12}_{i+1,i} &= \sin\theta\cos\psi\,|_{i+1,i}\\
\mathcal{R}^{13}_{i+1,i} &= -\sin\psi\cos\phi-\cos\theta\cos\psi\sin\phi\,|_{i+1,i}\\
\mathcal{R}^{21}_{i+1,i} &= -\sin\theta\cos\phi\,|_{i+1,i}\\
\mathcal{R}^{22}_{i+1,i} &= \cos\theta\,|_{i+1,i}\\
\mathcal{R}^{23}_{i+1,i} &= \sin\theta\sin\phi\,|_{i+1,i}\\
\mathcal{R}^{31}_{i+1,i} &= \cos\psi\sin\phi+\cos\theta\sin\psi\cos\phi\,|_{i+1,i}\\
\mathcal{R}^{32}_{i+1,i} &= \sin\theta\sin\psi\,|_{i+1,i}\\
\mathcal{R}^{33}_{i+1,i} &= \cos\psi\cos\phi-\cos\theta\sin\psi\sin\phi\,|_{i+1,i}.
\end{aligned} \tag{18}$$

Here the angular variables have the following ranges: For the inclination angle θ we take $\theta\in[0,\pi]$ mod(2π) and for the two azimuthal angles we choose $\phi\in[-\pi,\pi]$ mod(2π) and $\psi\in[-\pi,\pi]$ mod(2π). Note that since the angular variables are elements of the transfer matrix that takes the discrete Frenet frame from the vertex i to the vertex $i+1$, they are all to be interpreted as link variables that are defined on the bonds connecting the vertices.

From (15) we get the following condition:

$$\mathbf{b}_{i+1}\cdot\mathbf{t}_i=0.$$

Thus for each bond $(i,i+1)$

$$\sin\theta\sin\phi=0$$

and we conclude from (13)–(16) that for all i we must have

$$\phi_{i+1,i}=0.$$

This simplifies the discrete Frenet equation into

$$\begin{pmatrix}\mathbf{n}_{i+1}\\ \mathbf{b}_{i+1}\\ \mathbf{t}_{i+1}\end{pmatrix} = \begin{pmatrix}\cos\psi\cos\theta & \cos\psi\sin\theta & -\sin\psi\\ -\sin\theta & \cos\theta & 0\\ \sin\psi\cos\theta & \sin\psi\sin\theta & \cos\psi\end{pmatrix}_{i+1,i}\begin{pmatrix}\mathbf{n}_i\\ \mathbf{b}_i\\ \mathbf{t}_i\end{pmatrix}$$

$$\equiv \mathcal{R}_{i+1,i}\begin{pmatrix}\mathbf{n}_i\\ \mathbf{b}_i\\ \mathbf{t}_i\end{pmatrix}. \tag{19}$$

Here

$$\cos\psi_{i+1,i}=\mathbf{t}_{i+1}\cdot\mathbf{t}_i \tag{20}$$

is the discrete *bond angle* and

$$\cos\theta_{i+1,i}=\mathbf{b}_{i+1}\cdot\mathbf{b}_i \tag{21}$$

is the discrete *torsion angle*. Geometrically, the bond angle $\psi_{i+1,i}$ measures the angle between $\mathbf{t}_{i+1}$ and $\mathbf{t}_i$ around $\mathbf{b}_{i+1}$ on the plane that is determined by the three vertices (C_i,C_{i+1},C_{i+2}) (Fig. 5). The torsion angle $\theta_{i+1,i}$ measures the angle between the two planes that are determined by the vertices (C_{i-1},C_i,C_{i+1}) and (C_i,C_{i+1},C_{i+2}), respectively (Fig. 5).

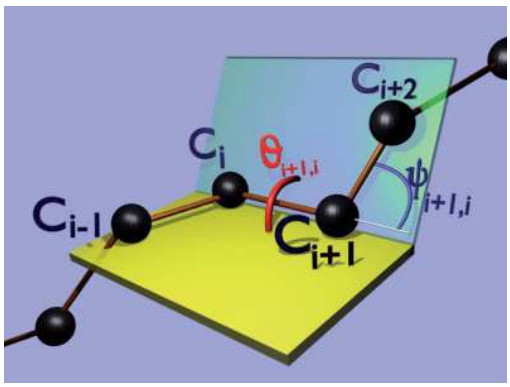

FIG. 5. (Color online) The bond angle $\psi_{i+1,i}$ is determined by the three vertices (C_{i-1},C_i,C_{i+1}). The torsion angle $\theta_{i+1,i}$ is the angle between the two planes determined by vertices (C_{i-1},C_i,C_{i+1}) and (C_i,C_{i+1},C_{i+2}).

We give these planes an orientation in $\mathbb{R}^3$ by extending the range of the torsion angle from $\theta_{i+1,i} \in [0,\pi]$ into $\theta_{i+1,i} \in [-\pi,\pi] \bmod(2\pi)$. This introduces a discrete $\mathbb{Z}_2$ symmetry

$$\mathbb{Z}_2 \ : \theta_{i+1,i} \leftrightarrow -\theta_{i+1,i} \tag{22}$$

that we find useful in the sequel.

We recall the Rodrigues formula

$$e^{\alpha\mathbb{U}} = \mathbb{I} + \mathbb{U}\sin\alpha + \mathbb{U}^2(1-\cos\alpha), \tag{23}$$

where

$$\mathbb{U} = \mathbf{u}\cdot\mathbf{T} \ = \ u^a T^a$$

and T^a are the SO(3) matrices and $\mathbf{u}$ is a unit vector. With these we can write the transfer matrix as follows:

$$\begin{aligned}\mathcal{R}_{i+1,i} &= \exp\{-\psi_{i+1,i}T^2\}\exp\{-\theta_{i+1,i}T^3\} \\ &= \ \exp\{-\alpha\mathbf{v}\cdot\mathbf{T}\}_{i+1,i},\end{aligned} \tag{24}$$

where

$$\alpha_{i+1,i} = 2\arccos\left[\frac{1}{4}(\mathbf{b}_{i+1}\cdot\mathbf{b}_i)(\mathbf{t}_{i+1}\cdot\mathbf{t}_i)\right]$$

and

$$\mathbf{v}_{i+1,i} = \frac{1}{\sin\frac{\alpha}{2}}\left(-\sin\frac{\psi}{2}\sin\frac{\theta}{2}\,,\ \sin\frac{\psi}{2}\cos\frac{\theta}{2}\,,\ \cos\frac{\psi}{2}\sin\frac{\theta}{2}\right)_{i+1,i}.$$

C. Gauge symmetries

Let us consider the effect of the discrete version of the local SO(2) rotation (4),

$$\begin{pmatrix}\mathbf{n}\\ \mathbf{b}\\ \mathbf{t}\end{pmatrix}_i \to e^{\Delta_i T^3}\begin{pmatrix}\mathbf{n}\\ \mathbf{b}\\ \mathbf{t}\end{pmatrix}_i. \tag{25}$$

For the covariance of the DF equation under (25) we need

$$e^{-\theta_{i+1,i}T^3} \to e^{\Delta_{i+1}T^3}e^{-\theta_{i+1,i}T^3}e^{-\Delta_i T^3} \tag{26}$$

$$e^{-\psi_{i+1,i}T^2} \to e^{\Delta_{i+1}T^3}e^{-\psi_{i+1,i}T^2}e^{-\Delta_{i+1}T^3}. \tag{27}$$

A direct computation shows that this implies the following transformation laws

$$\theta_{i+1,i} \to \theta_{i+1,i} + \Delta_i - \Delta_{i+1} \tag{28}$$

$$\psi_{i+1,i}T^2 \to \psi_{i+1,i}(T^2\cos\Delta_{i+1} - T^1\sin\Delta_{i+1}). \tag{29}$$

These are the discrete versions of the transformations of τ and κ in (6) and (7), respectively.

Explicitly, the gauge transformed transfer matrix is

$$\begin{aligned}
e^{\Delta_{i+1}T^3}\mathcal{R}_{i+1,i}\,e^{-\Delta_i T^3} &\equiv \mathcal{R}^{\Delta}_{i+1,i} \\
\mathcal{R}^{\Delta\ 11}_{i+1,i} &= \cos\Delta\cos\theta_\Delta\cos\psi + \sin\Delta\sin\theta_\Delta \\
\mathcal{R}^{\Delta\ 12}_{i+1,i} &= \cos\Delta\sin\theta_\Delta\cos\psi - \sin\Delta\cos\theta_\Delta \\
\mathcal{R}^{\Delta\ 13}_{i+1,i} &= -\cos\Delta\sin\psi \\
\mathcal{R}^{\Delta\ 21}_{i+1,i} &= \sin\Delta\cos\theta_\Delta\cos\psi - \cos\Delta\sin\theta_\Delta \\
\mathcal{R}^{\Delta\ 22}_{i+1,i} &= \sin\Delta\sin\theta_\Delta\cos\psi + \cos\Delta\cos\theta_\Delta \\
\mathcal{R}^{\Delta\ 23}_{i+1,i} &= -\sin\Delta\sin\psi \\
\mathcal{R}^{\Delta\ 31}_{i+1,i} &= \cos\theta_\Delta\sin\psi \\
\mathcal{R}^{\Delta 32}_{i+1,i} &= \sin\theta_\Delta\sin\psi \\
\mathcal{R}^{\Delta 33}_{i+1,i} &= \cos\psi.
\end{aligned} \tag{30}$$

We have here used the notation

$$\begin{aligned}\Delta &\equiv \Delta_{i+1} \\ \theta_\Delta &\equiv \theta_{i+1,i} + \Delta_i\end{aligned} \tag{31}$$

and the corresponding general frame Frenet equation is

$$\begin{pmatrix}\mathbf{e}_1\\ \mathbf{e}_2\\ \mathbf{t}\end{pmatrix}_{i+1} = \mathcal{R}^{\Delta}_{i+1,i}\begin{pmatrix}\mathbf{e}_1\\ \mathbf{e}_2\\ \mathbf{t}\end{pmatrix}_i. \tag{32}$$

Note that even though the explicit matrix elements in (30) do not have a manifestly covariant form in terms of the link variables, the gauge-transformed transfer matrix is by construction a covariant link variable.

D. Continuum limit

The different choices of Δ_i in (32) correspond to different generalized Frenet frames. We shall now verify that with the general version of transfer matrix (30), this indeed yields the generalized Frenet equation (5) in the continuum limit where the distances between the vertices C_i of the curve vanish, provided the limit is a class $\mathcal{C}^3$ curve.

$$|\mathbf{r}_{i+1} - \mathbf{r}_i| \approx \epsilon \to 0.$$

We define

$$\begin{aligned}\psi_{i+1,i} &= \epsilon\kappa_{i+1,i}, \\ \theta_{i+1,i} &= \epsilon\tau_{i+1,i}, \\ \Delta_{i+1} - \Delta_i &= \epsilon\sigma_{i+1,i}, \\ \tfrac{1}{2}(\Delta_{i+1} + \Delta_i) &= \eta_{i+1,i},\end{aligned} \tag{33}$$

where $\sigma_{i+i,i}$ are some finite constants. When we expand (32) in ϵ we get in the leading order

$$\begin{aligned}\frac{1}{\epsilon}\left[\begin{pmatrix}\mathbf{e}_1\\ \mathbf{e}_2\\ \mathbf{t}\end{pmatrix}_{i+1} - \begin{pmatrix}\mathbf{e}_1\\ \mathbf{e}_2\\ \mathbf{t}\end{pmatrix}_i\right] \\ = \begin{pmatrix}0 & (\tau-\sigma) & -\kappa\cos\eta\\ -(\tau-\sigma) & 0 & -\kappa\sin\eta\\ \kappa\cos\eta & \kappa\sin\eta & 0\end{pmatrix}_{i+1,i}\begin{pmatrix}\mathbf{e}_1\\ \mathbf{e}_2\\ \mathbf{t}\end{pmatrix}_i.\end{aligned} \tag{34}$$

If the $\epsilon \to 0$ exists it gives us the generalized continuum Frenet equation (5), with the identification

$$\sigma \to \dot{\eta}$$

and the identification (33) between the discrete torsion and curvature angles with their continuum counterparts.

E. Inflection points

Consider a piecewise linear curve that has a single isolated inflection point located at vertex C_i. A generalization to several

inflection points and straight segments is straightforward. By assumption, the preceding vertex C_{i-1} admits a Frenet frame. Since the tangent vectors $\mathbf{t}_i$ and $\mathbf{t}_{i-1}$ are parallel, at the vertex C_i both the normal vector $\mathbf{n}_i$ and the binormal vector $\mathbf{b}_i$ of a Frenet frame cannot be determined and the Frenet frame at C_i cannot be introduced. Consequently the torsion angle $\theta_{i,i-1}$ cannot be defined. But the definition of the bond angle involves only the tangent vectors so it can still be computed and from (20) we get

$$\psi_{i,i-1} = 0 \pmod{2\pi}.$$

In order to introduce a framing of the curve that covers the vertex C_i, we proceed as follows: We first deform the curve slightly by moving the vertex C_i in a direction of some arbitrarily chosen vector $\mathbf{u}$ that is not parallel with $\mathbf{t}_i$,

$$\mathbf{r}_i \to \mathbf{r}_i + \epsilon \cdot \mathbf{u}. \tag{35}$$

Here the limit $\epsilon \to 0$ is tacitly understood. The introduction of $\mathbf{u}$ removes the inflection point from the shifted vertex $\tilde{C}_i$ and this enables us to introduce a $\mathbf{u}$-dependent Frenet frame at the shifted vertex $\tilde{C}_i$. In the limit where ϵ vanishes we get a $\mathbf{u}$-dependent frame at the original vertex C_i, obtained by transferring the Frenet frame from the vertex C_{i-1} as follows,

$$\begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{t} \end{pmatrix}_i = \begin{pmatrix} \cos\hat{\theta} & \sin\hat{\theta} & 0 \\ -\sin\hat{\theta} & \cos\hat{\theta} & 0 \\ 0 & 0 & 1 \end{pmatrix}_{i,i-1} \begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \end{pmatrix}_{i-1}. \tag{36}$$

Here $\hat{\theta}_{i,i-1}$ is now description i.e., explicitly $\mathbf{u}$-dependent angle.

In order to establish that the frame can be chosen in a $\mathbf{u}$-independent manner we proceed to remove the explicit $\mathbf{u}$ dependence. For this we introduce the gauge transformation (28) in (36) which sends

$$\hat{\theta}_{i,i-1} \to \hat{\theta}_{i,i-1} + \Delta_{i-1} - \Delta_i.$$

Since we have the original Frenet frame at the vertex C_{i-1}, we also have

$$\Delta_{i-1} = 0.$$

But Δ_i is freely at our disposal and we may choose it so that any $\mathbf{u}$ dependence becomes removed. This leaves us with a $\mathbf{u}$-independent remainder that we may choose at our convenience,

$$\hat{\theta}_{i,i-1} - \Delta_i \equiv \hat{\Delta}_{i,i-1},$$

where $\hat{\Delta}_{i,i-1}$ is now by construction a $\mathbf{u}$-independent quantity at our disposal. Different choices correspond to different gauges.

Since $\mathbf{t}_i$ and $\mathbf{t}_{i+1}$ are not parallel, we can proceed to construct a frame at vertex C_{i+1} from the frame $(\mathbf{e}_1,\mathbf{e}_2,\mathbf{t})_i$ at vertex C_i using the transfer matrix (30). Since the remaining gauge parameters Δ_k with $k > i$ are all at our disposal, we may return to the Frenet frame or select any other convenient framing, at the vertex C_{i+1} and at all subsequent vertices. If the goal is to approximate a continuous space curve, in the limit of vanishing bond length the gauge parameters Δ_k should be selected in such a manner that in the continuum limit they yield the gauge function $\eta(s)$ and so the ensuing discrete transfer matrix smoothly goes over to its continuum limit (34).

F. Discrete gauge transformations

The transfer matrix $\mathcal{R}_{i+1,i}$ determines the curve in $\mathbb{R}^3$ up to rigid Galilean motions, i.e., global translations and spatial rotations. The improper spatial rotation group O(3) acts on each of the vertices $\mathbf{r}_k$ in (14) by a rotation matrix $\mathcal{O} \in \mathrm{O}(3)$ that sends each of the $\mathbf{r}_k$ into

$$\mathbf{r}_k \to \mathcal{O}\mathbf{r}_k.$$

As a consequence only the global orientation of the curve in $\mathbb{R}^3$ changes. An example is the improper rotation that inverts the curve in $\mathbb{R}^3$ by reversing the direction of each tangent vector

$$\mathbf{t}_i \to -\mathbf{t}_i$$

but with no effect on the $\mathbf{n}_i$ and $\mathbf{b}_i$. From the explicit form of the transfer matrix in (19) we conclude that this corresponds to the following global version of (28) and (29)

$$\theta_i \to \theta_i$$
$$\psi_i \to -\psi_i.$$

That is, $\Delta_i = \pi$ for all i. Consequently, if we include this improper rotation in our gauge structure we can restrict the range of ψ_i from $\psi_i \in [-\pi,\pi] \bmod(2\pi)$ to $\psi_i \in [0,\pi] \bmod(2\pi)$, but we prefer to continue with the extended range.

Similarly, we can introduce the improper rotation that sends

$$\mathbf{b}_i \to -\mathbf{b}_i$$

with no effect on $\mathbf{t}_i$ and $\mathbf{n}_i$. Since the $\mathbf{t}_i$ remain intact, the curve does not change, and from the DF equation (19) we conclude that this corresponds to the following global $\mathbb{Z}_2$ transformation:

$$\theta_i \to -\theta_i$$
$$\psi_i \to \psi_i.$$

This is the $\mathbb{Z}_2$ symmetry that we have introduced in (22) to extend the range of θ_i from $\theta_i \in [0,\pi]$ to $\theta_i \in [-\pi,\pi] \bmod(2\pi)$. We note that this symmetry of the underlying curve can not be reproduced by the gauge transformation (28) and (29); nevertheless, the curve remains intact since the $\mathbf{t}_i$ do not change.

Another useful discrete transformation in our subsequent discrete curve analysis is the proper rotation that at a given vertex C_i sends

$$\mathbf{b}_i \to -\mathbf{b}_i$$
$$\mathbf{n}_i \to -\mathbf{n}_i$$

but with no effect on $\mathbf{t}_i$ so the curve remains intact. This rotation is obtained by selecting $\Delta_{i+1} = \pi$ and with all $\Delta_k = 0$ at the preceding vertices C_k (with $k \leqslant i$). Since the Δ_{i+1} appears in the gauge transformation law of both $\theta_{i+1,i}$ and $\theta_{i+2,i+1}$, this leads to the following realization of the gauge transformation (28) and (29)

$$\theta_{i+1,i} \to \theta_{i+1,i} - \pi$$
$$\theta_{i+2,i+1} \to \theta_{i+1,i} + \pi$$

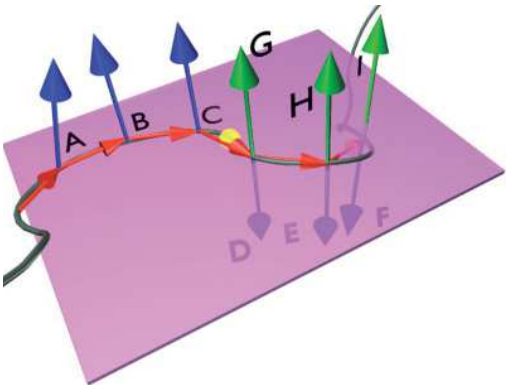

FIG. 6. (Color online) A continuous plane curve with an inflection point such as the one in Fig. 1, together with its discrete approximation. The tangent vectors $\mathbf{t}_i$ of the discrete approximation can be chosen so two neighbors are never parallel and thus a discrete Frenet frame can be introduced at each vertex. When we pass through the inflection point the direction of the binormal vectors following (**A**,**B**,**C**) becomes reflected in the plane into (**D**,**E**,**F**) and there is a discontinuity in the Frenet framing. But if we introduce the gauge transformation (37) at vertices after the inflection point, the ensuing framing (**A**,**B**,**C**,**G**,**H**,**I**) is continuous.

$$\psi_{i+1,i} \to -\psi_{i+1,i}.$$

If we generalize this gauge transformation by selecting

$$\Delta_k = \pi \quad \text{for } k \geqslant i+1$$

with

$$\Delta_k = 0 \quad \text{for } k < i+1,$$

where the vertex C_i is preselected, the gauge transformation becomes

$$\begin{aligned} \theta_{i+1,i} &\to \theta_{i+1,i} - \pi \\ \psi_{k+1,k} &\to -\psi_{k+1,k} \quad \text{for all } k \geqslant i. \end{aligned} \tag{37}$$

Since the bond angle is the discrete version of the Frenet curvature (33), we recognize here the discrete analog of the continuum gauge transformation (10) and (11). For a piecewise linear discretization of a plane curve such as the one Fig. 1, this enables us to introduce a framing that captures the kink-soliton behavior (1) and (12) of the inflection point, with the change of sign in curvature at the soliton position (Fig. 6).

G. Curve construction

An example of problems where the present formalism can be applied is the construction of a discrete and piecewise linear curve from the known values of its bond and torsion angles. These angles can be constructed, for example, using an energy principle to locate a minimum energy configuration of some energy functional

$$E(\psi_{k+1,k},\theta_{k+1,k}).$$

We may define the angles using the Frenet frame. Examples of energy functionals have been discussed in Refs. [3,12].

Three vertices are needed to specify the position and the overall rotational orientation of the curve. To compute a single bond angle from the curve, we need three vertices while for the torsion angle we need four; see Fig. 5. Consequently, from the first three initial positions of the curve, $(\mathbf{r}_0,\mathbf{r}_1,\mathbf{r}_2)$, we can compute the first bond angle $\psi_{1,0}$. But in order to compute the first pair $(\psi_{2,1},\theta_{2,1})$ we also need to specify $\mathbf{r}_3$.

Here we are interested in the inverse problem where the set of angles $\{\psi_{k+1,k},\theta_{k+1,k}\}$ are assumed to be known. Depending on the boundary conditions for the energy functional, the known initial data may also include numerical values of $(\psi_{1,0},\theta_{1,0})$, even though $\theta_{1,0}$ lacks a geometric interpretation. In such a case we can immediately proceed to the computation of the entire curve using (19) or, alternatively, using the transfer matrix (32), starting from an initial choice of frame $(\mathbf{n}_0,\mathbf{b}_0,\mathbf{t}_0)$. Different initial choices are related to each other by a *global*, i.e., index i-independent, parameter Δ in (28) and (29). We get both the frame at the vertex k and its location $\mathbf{r}_k$ when we also employ (14), starting from a given initial value $\mathbf{r}_0(=0)$.

In general we expect to have a situation where the three first points $(\mathbf{r}_0,\mathbf{r}_1,\mathbf{r}_2)$ are given. From these points we get the two tangent vectors $\mathbf{t}_0$ and $\mathbf{t}_1$. We then use (15) and (16) to complete the Frenet frame at the location $\mathbf{r}_1$. We identify the bond angle $\psi_{1,0}$ with the angle between the two vectors $\mathbf{t}_0$ and $\mathbf{t}_1$ using (20). This bond angle may or may not be determined by the energy functional. If it is determined, the angle between $\mathbf{t}_0$ and $\mathbf{t}_1$ is determined and instead of fully specifying $\mathbf{r}_2$ we only need to specify its distance from $\mathbf{r}_1$ and the remaining directional angle that we may call $\theta_{1,0}$.

For a practical algorithmic implementation the following choice can be convenient,

$$\begin{aligned} \mathbf{r}_0 &= \delta_{1,0}\begin{pmatrix} -\cos\psi_{1,0} \\ \sin\psi_{1,0} \\ 0 \end{pmatrix} \\ \mathbf{r}_1 &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \\ \mathbf{t}_0 &= \begin{pmatrix} \cos\psi_{1,0} \\ -\sin\psi_{1,0} \\ 0 \end{pmatrix} \end{aligned}$$

$$\mathbf{n}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \mathbf{b}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \mathbf{t}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \tag{38}$$

where we have introduced the notation

$$\delta_{k+1,k} = |\mathbf{r}_{k+1} - \mathbf{r}_k|$$

for the segment lengths. The generalized Frenet frame together with the corresponding location of the vertex $\mathbf{r}_{i+1}$ can then be computed by iterative application of

$$\begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \\ \mathbf{r} \end{pmatrix}_{i+1} = \mathcal{T}_{i+1,i}\begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \\ \mathbf{r} \end{pmatrix}_{i} = \begin{pmatrix} & (\mathcal{R}) & & 0 \\ & & & 0 \\ & & & 0 \\ 0 & 0 & \delta & 1 \end{pmatrix}_{i+1,i}\begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \\ \mathbf{r} \end{pmatrix}_{i+1}. \tag{39}$$

This can be directly generalized into

$$\begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \\ \mathbf{r} \end{pmatrix}_{i+1} = \mathcal{T}^{\Delta}_{i+1,i} \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \\ \mathbf{r} \end{pmatrix}_{i}$$

$$= \begin{pmatrix} & & & 0 \\ & (\mathcal{R}^{\Delta}) & & 0 \\ & & & 0 \\ \delta_1 & \delta_2 & \delta_3 & 1 \end{pmatrix}_{i+1,i} \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \\ \mathbf{r} \end{pmatrix}_{i+1},$$

where $\mathcal{R}^{\Delta}$ is the matrix (30) and the $\delta_1,\delta_2,\delta_3$ are the components of the vector

$$\vec{\delta}_{k+1,k} = \delta_{k+1,k} \begin{pmatrix} \cos\alpha \sin\beta \\ \sin\alpha \sin\beta \\ \cos\beta \end{pmatrix}_{k+1,k}.$$

When $\beta = 0$ (and $\Delta = 0$) we obtain the transfer matrix (39) with $\mathbf{t}_k$ the tangent vector of the curve, while for general (α,β) the tangent of the curve is in the direction of $\vec{\delta}$ in the $(\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3)$ frame. Thus this transfer matrix provides a rule for transporting an *a priori* arbitrarily oriented orthogonal frame along the curve.

Of particular interest is the construction of a discrete version of Bishop's parallel transport frame [7] as a gauge transformed version of the discrete Frenet frame. Since the Frenet frame starts with $(\psi_{2,1},\theta_{2,1})$ and can be constructed once $(\mathbf{r}_0,\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3)$ are known (unless we introduce $\theta_{1,0}$ which lacks a geometric interpretation), we assume this to be the case. The discrete version of Bishop's frame is obtained by gauge transformation from the Frenet frame by demanding that

$$\theta_{2,1} \to \theta_{2,1} + \Delta_1 - \Delta_2 = 0.$$

We can freely choose

$$\Delta_1 = 0$$

as an initial condition, and, consequently, we arrive at Bishop's frame by selecting

$$\Delta_2 = \theta_{2,1}.$$

For Δ_3 we get similarly from

$$\theta_{3,2} \to \theta_{3,2} + \Delta_2 - \Delta_3 = 0$$

that

$$\Delta_3 = \theta_{2,1} + \theta_{3,2}$$

and thus the discrete version of Bishop's parallel transport frame is related to the discrete Frenet frame by gauge transformations

$$\Delta_k = \sum_{i=1}^{k-1} \theta_{i+1,i}.$$

When we substitute this in (30) with (31), we find that the transfer matrix (30) simplifies into

$$\mathcal{R}^{B\ 11}_{i+1,i} = 1 + \cos^2\Theta_{\Delta}(\cos\psi_{i+1,i} - 1)$$

$$\begin{aligned}
\mathcal{R}^{B\ 12}_{i+1,i} &= \sin\Theta_{\Delta}\cos\Theta_{\Delta}(\cos\psi_{i+1,i} - 1) \\
\mathcal{R}^{B\ 13}_{i+1,i} &= -\cos\Theta_{\Delta}\sin\psi_{i+1,i} \\
\mathcal{R}^{B\ 21}_{i+1,i} &= \sin\Theta_{\Delta}\cos\Theta_{\Delta}(\cos\psi_{i+1,i} - 1) \\
\mathcal{R}^{B\ 22}_{i+1,i} &= 1 + \sin^2\Theta_{\Delta}(\cos\psi_{i+1,i} - 1) \\
\mathcal{R}^{B\ 23}_{i+1,i} &= -\sin\Theta_{\Delta}\sin\psi_{i+1,i} \\
\mathcal{R}^{B\ 31}_{i+1,i} &= \cos\Theta_{\Delta}\ \sin\psi_{i+1,i} \\
\mathcal{R}^{B\ 32}_{i+1,i} &= \sin\Theta_{\Delta}\ \sin\psi_{i+1,i} \\
\mathcal{R}^{B\ 33}_{i+1,i} &= \cos\psi, \qquad (40)
\end{aligned}$$

where now

$$\Theta_{\Delta} \equiv \sum_{k=1}^{i} \theta_{k+1,k}$$

and with (32), we can construct the discrete version of Bishop's parallel transport frame at each vertex C_i.

IV. FRAMING OF FOLDED PROTEINS

As an application we utilize the DF equation to investigate the framing of the folded proteins in the PDB [9]. We are particularly interested in the existence and characterization of a *preferred* framing that derives and directly reflects the physical properties of the folded proteins. The identification of such a preferred framing, if it exists, should help to pinpoint the physical principles that determine how proteins fold.

From the PDB we get the three-dimensional coordinates of all the different atoms in a folded protein. The overall fold geometry is described by the location of the central C_α carbons that determine the protein backbone. We take the C_α carbons to be the vertices in a discrete and piecewise linear curve that models the backbone. We then use the C_α coordinates to compute the corresponding Frenet framing. For this we first apply (13), (15), and (16) to obtain the orthonormal basis vectors at each vertex. We then construct the transfer matrices by evaluating the bond and torsion angles from (20) and (21).

A. $\mathbb{Z}_2$ Frenet framing and solitons

We start by analyzing in detail an explicit example, the chicken villin headpiece subdomain HP35 (PDB code 1YRF [9]). This is a naturally existing 35-residue protein, with three α helices separated from each other by two loops. This protein continues to be the subject to very extensive studies both experimentally [14–17] and theoretically [18–22]. We note that the overall resolution in the experimental x-ray data in PDB is 1.07Å in root-mean-square deviation [16].

We first compute the backbone Frenet frame bond and torsion angles $(\psi_{i+1,i},\theta_{i+1,i})$ from the PDB coordinates of the HP35 C_α carbons. The result is shown in Fig. 7 (left).

We inquire whether the loop regions contain inflection points. As we have previously explained, for example, in connection with Fig. 6, the inflection points can be difficult to identify in terms of the bond angles of the discrete Frenet framing alone. But as is apparent from Fig. 6, we can expect that an inflection point is located in the vicinity of vertices where the Frenet frame torsion angle is subject to strong local fluctuations. Thus we proceed to inspect the data in Fig. 7

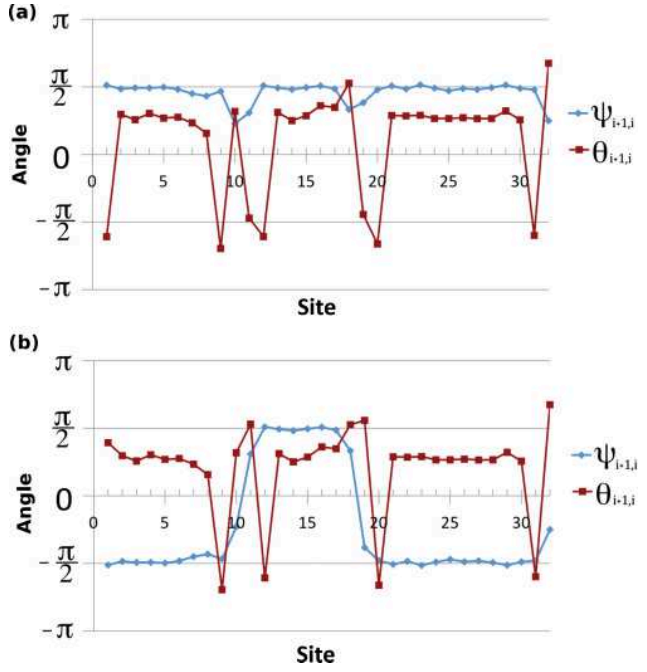

FIG. 7. (Color online) (Top) The Frenet frame bond angle (Ψ) and torsion angle (Θ) along the HP35 backbone. In this frame the potential presence of an inflection point is visible only in large local variations of torsion angle. (Bottom) The outcome of $\mathbb{Z}_2$ gauge transformations (37) at the loop regions. The result clearly reveals the presence of inflection points, and they are located between the sites where the (gauge transformed) bond angle changes its sign. This can also be used to identify the center of the loop. Note how closely the profile of the bond angle in the bottom figure resembles that of the kink-soliton in the right-hand side of Fig. 2.

(left) using the gauge transformation (37) to scrutinize the loop regions where the Frenet torsion angle in Fig. 7 (left) is strongly fluctuating. This leads us to a particular version of the $\mathbb{Z}_2$ Frenet frame, with bond and torsion angles as in Fig. 7 (right).

By comparing the bond angles in Fig. 7 (right) with the kink-soliton profile in the right hand side of Fig. 2 we observe that the bond angles of our gauge transformed frames at each of the loops have assumed the distinctive hallmark profile of a (discrete) kink-soliton that interpolates between the adjacent α helices. In particular, we can unambiguously pinpoint the centers of the loops to the locations of the inflection points on the curve: The inflection points are between the vertices where the bond angle in our gauge transformed frame changes its sign.

We have performed a similar analysis to several proteins in the PDB, and some of our results where the techniques of the present article are utilized have been reported in Refs. [4–6]. The results are remarkably consistent: In every secondary superstructure that we have studied where a loop connects two α helices and/or β strands, after appropriate $\mathbb{Z}_2$ gauge transformations the profile of the bond angles in the loop can be described with sub-Ångström accuracy in terms of a discrete version of the kink-soliton in Fig. 2. The two asymptotic ground states at $s = \pm c$ in this figure correspond to the α helices and/or β strands at the ends of the loop. For the α helices we have the Frenet frame values very close to

$$(\psi,\theta)_\alpha \approx (1.57,0.87) \sim \left(\frac{\pi}{2},1\right).$$

The β strands can also be interpreted as helices but in the "collapsed" limit with the approximative values

$$(\psi,\theta)_\beta \approx (\pm 1.0, -2.9) \sim (\pm 1, -\pi).$$

Consequently, it appears that these α-helix/β-strand-loop-α-helix/β-strand superstructures are indeed *inflection point solitons* with the qualitative profile of (1). We remark that a long loop may also consist of a number of inflection points, i.e., it can be a multisoliton configuration.

B. Physics-based framing

In every amino acid except glycine, there is a C_β carbon that is covalently bonded to a C_α carbon. The positioning of these C_β carbons in relation to their C_α carbons characterizes the relative orientation of the amino acid side chains along the protein backbone and can be used to introduce a distinctive framing of the backbone; the case of glycine can be treated like that of an inflection point. Since the interactions between different amino acids are presumed to have a pivotal role both during the folding process and in the stabilization of the native fold, the C_β framing should be a natural choice to intimately reflect the physical principles that determine the fold geometry of the backbone. Consequently, one way to try and understand the physical principles that determine how a protein folds could be to investigate the C_β framing along the protein backbone. Here we propose that a practical approach is to look for gauge parameters (25) that relate the C_β frames to some purely geometrically determined frames such as the Frenet frames or parallel transport frames. The identification of the rules that determine the relevant gauge parameters Δ_i should then provide insight to the physical principles that underlie the protein-folding phenomenon.

The C_β framing is constructed from the tangent vectors $\mathbf{t}$ of the backbone and the unit vectors $\mathbf{c}$ that point from the C_α carbons toward their C_β carbon. The framing is obtained by Gram-Schmidt algorithm by first introducing the unit vector

$$\mathbf{p} = \frac{\mathbf{t}\times\mathbf{c}}{||\mathbf{t}\times\mathbf{c}||}$$

and then completing it into an orthonormal frame $(\mathbf{t},\mathbf{p},\mathbf{q})$ at each C_α vertex, where

$$\mathbf{q} = \mathbf{t}\times\mathbf{p}.$$

In order to characterize the rules that determine the gauge parameter Δ_i relating a C_β frame to the corresponding Frenet frame, we have investigated the statistical distribution of the $\mathbf{c}_i$ vectors in the PDB proteins in the Frenet framing of the backbone. For this we introduce, at each backbone vertex, the inclination angle $\chi_i \in [0,\pi]$ between the tangent vector $\mathbf{t}_i$ and the corresponding vector $\mathbf{c}_i$, together with the azimuthal angle $\varphi_i \in [-\pi,\pi]$ between the normal vector $\mathbf{n}_i$ and the projection of $\mathbf{c}_i$ on the $(\mathbf{n}_i,\mathbf{b}_i)$ plane; see Fig. 8.

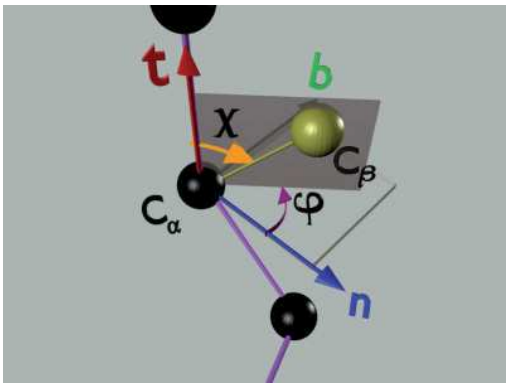

FIG. 8. (Color online) The definition of the angles χ_i and φ_i that describe the location of the ith C_β carbon with respect to the Frenet frame along the C_α backbone. The distance between the C_α and C_β carbons is within the range of 1.56–1.57 Å.

We first consider the C_β framing of the HP35. When we compute the directions of the individual vectors $\mathbf{c}_i$ in the Frenet frame, we get the result that we display in Fig. 9.

Remarkably, the directions of the $\mathbf{c}_i$ vectors in the Frenet frame are *relatively* site independent. This implies that at least in the case of HP35, the parameters Δ_i that relate the C_β frame to the Frenet frame can be assigned to a high accuracy a constant and site independent value: The physically determined orthonormalized C_β frame appears to differ from the purely geometrically determined Frenet frame only by small nutations in the direction of the vectors $\mathbf{c}$ in the Frenet frame. We observe that these nutations are somewhat smaller in the helix regions than in the loops. We conclude that since the Frenet framing of HP35 is determined entirely by the backbone geometry so, too, are the orientations of the amino acids, with a surprisingly good accuracy.

In the general case, we have inspected the correlation between the C_β framing and the Frenet framing by performing

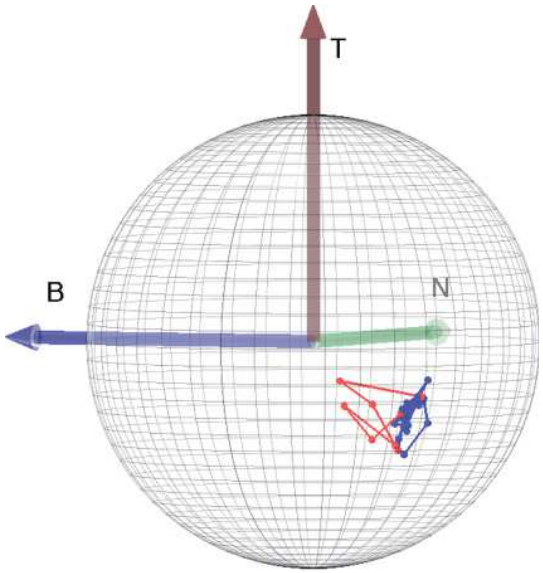

FIG. 9. (Color online) The nutation in the direction of the vectors $\mathbf{c}_i$ in the Frenet frame along 1YRF backbone.

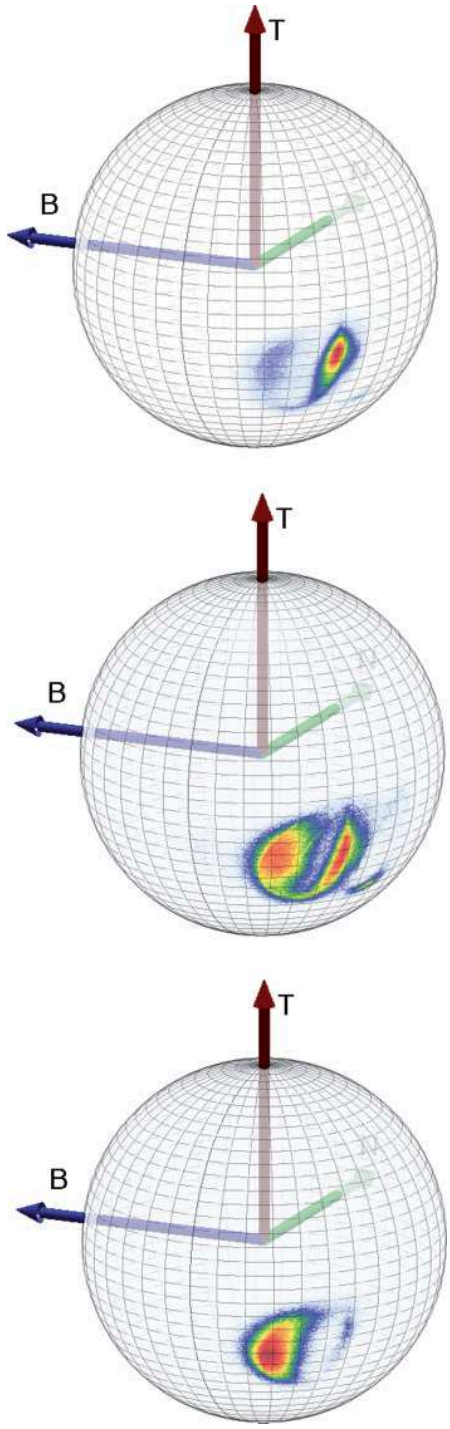

FIG. 10. (Color online) Kent plots of the C_β carbon vectors $\mathbf{c}$ for all sites of all proteins presently in PDB, with intensity proportional to the number of vectors. For α helices (top), the direction of $\mathbf{c}$ nutates very little around the direction $(\chi,\varphi) \approx (1.84, -2.20)$. For β strands (bottom) the nutation is somewhat more spread out, but still very clearly concentrated around $(\chi,\varphi) \approx (1.96, -2.47)$. Finally, for loops (middle) we observe the formation of a narrow arc that connects the α and β directions.

a statistical analysis of the directional distribution of the **c** vectors in the backbone Frenet frames for all amino acids in PDB. Our results are summarized in Fig. 10.

Where we display the statistical distribution of the angles (φ_i, χ_i) that we have defined in Fig. 8. We have used the PDB definition to identify the three structures we display separately (α helix, β strand, loop) but we note that there are sometimes ambiguities in determining whether a particular amino acid belongs to a α helix, β strand or a loop in particular when the amino acid is located in the vicinity of the border between these three classes.

We find that the observation we have made in the case of HP35 persists: The orientations of the C_β carbons in the Frenet frames are quite inert and essentially protein and amino acid independent. There is only a slight nutation around the statistical average value. Furthermore, the directions for the α helices and β strands are also almost the same, the difference in the statistical average is surprisingly small but nevertheless noticeable. In the case of loops, we find that the statistical distribution of the vector **c** in the Frenet frame displays a thin band that connects the α helices and β strands. This universality is somewhat unexpected, since only a small proportion of the loops connect an α helix with a β strand.

The overlapping regions between the three different classes in the Kent plots of Fig. 10 can be at least partly explained by the uncertainty in classifying amino acids in the vicinity of the border regions. We expect that a careful scrutiny of the class assignments of these amino acids will sharpen our statistical results. Alternatively, our approach could be developed into a technique to determine a more definite classification of those amino acids that are located in the border regions separating

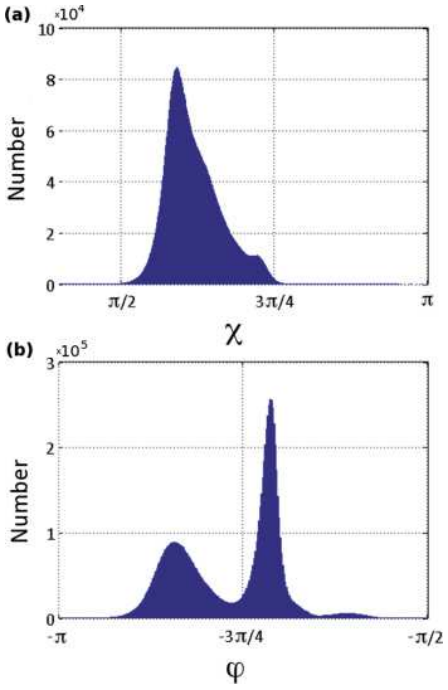

FIG. 11. (Color online) Frenet frame histogram of the distribution of (χ,φ) angles displayed in Fig. 10 for all C_β in the PDB. The histogram shows how the directions are subject to only very small deviations around their average values.

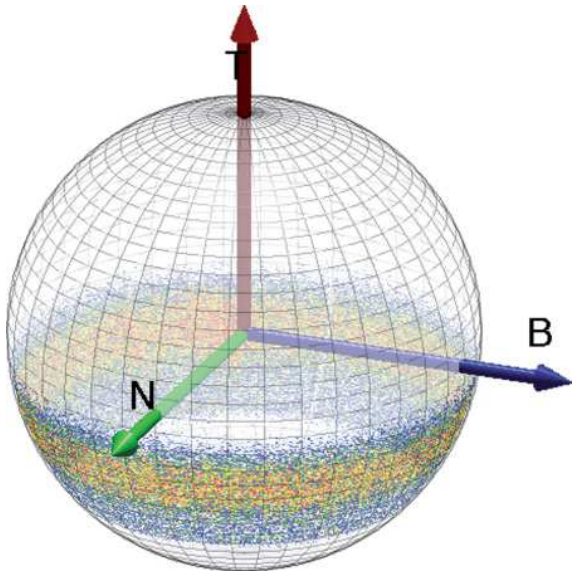

FIG. 12. (Color online) The same as in Fig. 10 but for all proteins in PDB using Bishop's parallel transport frame. In this frame the directions of the C_β carbons are distributed in a longitudinally uniform manner inside a segment of the Kent sphere.

the α helices, β strands, and the loops from each other. But even at this level of classifying the amino acids the results of our analysis imply that almost independently of the protein, when we traverse its backbone by orienting the camera gaze direction so it remains fixed in the Frenet frames, the directions of the C_β carbons are subject to only small nutations.

In Fig. 11 we display the histograms for the components of the C_β vectors $\mathbf{c}_i$ in terms of the χ and φ angles defined in Fig. 8. These histograms confirm that the directional variations of the $\mathbf{c}_i$ are surprisingly inert.

Finally, we have found that in Bishop's parallel transport frame the direction of the C_β carbon does not lead to such a regular structure formation as in the Frenet frame; see Fig. 12 where we plot the statistical distributions of the vectors **c** in the Bishop's frames.

V. CONCLUSIONS

We have scrutinized the problem of frame determination along piecewise linear discrete curves, including those with inflection points. Our approach is based on the transfer matrix method that has been previously applied extensively to investigate discrete integrable systems and lattice field theories. The introduction of a transfer matrix enables us to describe a framing in a covariant manner, with different frames related to each other by SO(2) gauge transformations that correspond to rotations in the normal planes of the curve. In particular our construction is not based on, and does not involve, any discretization of a continuum equation. Consequently, we can effortlessly describe curves that become fractals in the limit where the lattice spacing (e.g., the length of line segments) vanishes. But we have also verified that if the continuum limit exists as a class $\mathcal{C}^3$ differentiable curve, we arrive at the generalized version of the continuum Frenet equation. Furthermore, the manifest covariance of our

formalism under frame rotations enables us to investigate the framing of a physically determined discrete curve in a manner where the framing is based on and captures the properties of the underlying physical system. Consequently, we expect that our formalism has wide applications to the visualization of discrete curves and the determination of camera gaze positions in a variety of scenarios.

One notable outcome of our analysis is the identification of inflection points with the centers of loops, and the interpretation of loops as kink-solitons. In Refs. [3,12] we have already applied this identification to develop an ansatz based on (1) to succesfully describe the native folds of PDB proteins in terms of elementary functions.

As an example, we have investigated the framing of folded proteins in the Protein Data Bank. In this case no *valable* continuum description exist, due to the fact that the universality class of folded proteins is characterized by the presence of a nontrivial Hausdorff dimension. Consequently, any framing of folded proteins should be inherently discrete. In order to introduce a framing that directly relates to the physical properties of a folded protein, we have employed the relative orientation of the C_β carbons in the amino acids with respect to the ensuing backbone central C_α carbons. We have statistically analyzed the relative orientation of these C_β frames to the geometrically determined Frenet frames of the PDB proteins. We have found that the two framings are almost identical, and they differ from each other only by a practically amino acid independent global frame rotation: For the α helices the nutation in the orientation of the C_β carbons in the Frenet frame is *very* sharply concentrated around its statistically determined average direction. For β strands the result is very similar, with only a relatively small increase in the amplitude of nutations. Finally, in the case of loops we find that the orientation of the C_β carbons oscillates along a narrow circular arc that connects the α helices and β carbons. In each case, we have used the definition employed in the Protein Data Bank to identify the helix or loop class of the amino acid, and we note that the existing criteria for determining this class in the case of an amino acid that is located in the vicinity of the terminals of each structure is subject to interpretations. Consequently, we propose that there are several borderline cases that interfere destructively with the accuracy of our statistically determined results. We hope that our framing technique will eventually provide a refinement of the existing classification principles. The biophysical interpretation and biological relevance of our observations will be reported elsewhere.

ACKNOWLEDGMENTS

We thank Maxim Chernodub for many valuable discussions and Jack Quine for communications.

[1] A. J. Hanson, *Visualizing Quaternions* (Morgan Kaufmann Elsevier, London, 2006).
[2] J. B. Kuipers, *Quaternions and Rotation Sequences: A Primer with Applications to Orbits, Aerospace, and Virtual Reality* (Princeton University Press, Princeton, NJ, 1999).
[3] U. H. Danielsson, M. Lundgren, and A. J. Niemi, Phys. Rev. E **82**, 021910 (2010).
[4] M. Chernodub, S. Hu, and A. J. Niemi, Phys. Rev. E **82**, 011916 (2010).
[5] N. Molkenthin, S. Hu, and A. J. Niemi, Phys. Rev. Lett. **106**, 078102 (2011).
[6] S. Hu, A. Krokhotin, A. J. Niemi, and X. Peng, Phys. Rev. E **83**, 041907 (2011).
[7] R. L. Bishop, Am. Math. Mon. **82**, 246 (1974).
[8] I. Montvay and G. Münster, *Quantum Fields on a Lattice* (Cambridge University Press, Cambridge, 1994).
[9] H. M. Berman, K. Henrick, H. Nakamura, and J. L. Markley, Nucleic Acids Res. **35**, D301 (2007)
[10] N. Manton and P. Sutcliffe, *Topological Solitons* (Cambridge University Press, Cambridge, 2004).
[11] M. Spivak, *A Comprehensive Introduction to Differential Geometry* (Publish or Perish, Houston, 1979), Vol. 5.
[12] A. J. Niemi, Phys. Rev. D **67**, 106004 (2003).
[13] H. Hasimoto, J. Fluid Mech. **51**, 477 (1972).
[14] C. J. McKnight, P. T. Matsudaira, and P. S. Kim, Nat. Struct. Biol. **4**, 180 (1997).
[15] J. Meng, D. Vardar, Y. Wang, H. C. Guo, J. F. Head, and C. J. McKnight, Biochemistry **44**, 11963 (2005).
[16] T. K. Chiu, J. Kubelka, R. Herbst Irmer, W. A. Eaton, J. Hofrichter, and D. R. Davies, Proc. Natl. Acad. Sci. U.S.A. **102**, 7517 (2005).
[17] L. Wickstrom, Y. Bi, V. Hornak, D. P. Raleigh, and C. Simmerling, Biochemistry **46**, 3624 (2007).
[18] G. Jayachandran, V. Vishal, and V. S. Pande, J. Chem. Phys. **124**, 164902 (2006).
[19] D. L. Ensign, P. M. Kasson, and V. S. Pande, J. Mol. Biol. **374**, 806 (2007).
[20] H. Lei and Y. Duan, J. Mol. Biol. **370**, 196 (2007).
[21] P. L. Freddolino and K. Schulten, Biophys. J. **97**, 2338 (2009).
[22] P. L. Freddolino, C. B. Harrison, Y. Liu, and K. Schulten, Nat. Phys. **6**, 751 (2010).

ERRATA : Discrete Frenet frame, inflection point solitons, and curve visualization with applications to folded proteins [Phys. Rev. E 83, 061908 (2011)]

Shuangwei Hu, Martin Lundgren, and Antti J. Niemi

The two symbols "$\cdot\ c$" in FIG. 2 are mistakenly shown and should be "$-c$" instead. The last line of Eq. (38) on page 061908-8 should read

$$\mathbf{t}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \ \mathbf{n}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \ \mathbf{b}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

.

Chapitre 7

Topological solitons and folded proteins

Maxim Chernodub, Shuangwei Hu, and Antti J. Niemi

Physical Review E 82, 011916 (2010)

Topological solitons and folded proteins

Maxim Chernodub,[1,2,*] Shuangwei Hu,[1,3,†] and Antti J. Niemi[1,3,‡]
[1]*Laboratoire de Mathematiques et Physique Theorique CNRS UMR 6083, Fédération Denis Poisson, Université de Tours, Parc de Grandmont, F37200 Tours, France*
[2]*Department of Mathematical Physics and Astronomy, Ghent University, Krijgslaan 281, 59, Gent B-9000, Belgium*
[3]*Department of Physics and Astronomy, Uppsala University, P.O. Box 803, S-75108 Uppsala, Sweden*
(Received 1 April 2010; published 21 July 2010)

We argue that protein loops can be described by topological domain-wall solitons that interpolate between ground states which are the α helices and β strands. We present an energy function that realizes loops as soliton solutions to its equation of motion, and apply these solitons to model a number of biologically active proteins including 1VII, 2RB8, and 3EBX (Protein Data Bank codes). In all the examples that we have considered we are able to numerically construct soliton solutions that reproduce secondary structural motifs such as α-helix-loop-α-helix and β-sheet-loop-β-sheet with an overall root-mean-square-distance accuracy of around 1.0 Å or less for the central α-carbons, i.e., close to the limits of current experimental accuracy.

DOI: 10.1103/PhysRevE.82.011916 PACS number(s): 87.15.Cc, 05.45.Yv

Solitons are ubiquitous and widely studied objects that can be materialized in a variety of practical and theoretical scenarios [1,2]. For example solitons can be deployed for data transmission in transoceanic cables, for conducting electricity in organic polymers [1], and they may also transport chemical energy in proteins [3]. Solitons explain the Meissner effect in superconductivity and dislocations in liquid crystals [1]. They also model hadronic particles, cosmic strings, and magnetic monopoles in high energy physics [2] and so on. The first soliton to be identified is the *Wave of Translation* that was observed by John Scott Russell in the Union Canal of Scotland. This wave can be accurately described by an exact soliton solution of the Korteweg-de Vries (KdV) equation [1]. At least in principle it can also be constructed in an atomary level simulation where one accounts for each and every water molecule in the Canal, together with all of their mutual interactions. However, in such a *Gedanken* simulation it would probably become a real challenge to unravel the collective excitations that combine into the *Wave of Translation* without any guidance from the known soliton solution of the KdV equation: Solitons can *not* be constructed simply by adding up small perturbations around some ground state. Instead, a (topological) soliton emerges when non-linear interactions combine elementary constituents into a localized collective excitation that is stable against small perturbations and cannot decay, unwrap or disentangle [1,2].

In this Communication we argue that topological solitons describe proteins in their native folded state [4,6]. We characterize a folded protein by the Cartesian coordinates $\mathbf{r}_i$ of its N central α carbons, with $i=1,\ldots,N$. For many biologically active proteins these coordinates can be downloaded from protein data bank (PDB) [7]. Alternatively, the protein can be described in terms of its bond and torsion angles that can be computed from the PDB data. For this we introduce the tangent vector $\mathbf{t}_i$ and the binormal vector $\mathbf{b}_i$

$$\mathbf{t}_i = \frac{\mathbf{r}_{i+1}-\mathbf{r}_i}{|\mathbf{r}_{i+1}-\mathbf{r}_i|} \quad \& \quad \mathbf{b}_i = \frac{\mathbf{t}_{i-1}\times\mathbf{t}_i}{|\mathbf{t}_{i-1}-\mathbf{t}_i|}. \tag{1}$$

Together with the normal vector $\mathbf{n}_i=\mathbf{b}_i\times\mathbf{t}_i$ we then have three vectors that are subject to the discrete Frenet equation [8].

$$\begin{pmatrix}\mathbf{n}_{i+1}\\ \mathbf{b}_{i+1}\\ \mathbf{t}_{i+1}\end{pmatrix} = \exp\{-\kappa_i\cdot T^2\}\cdot\exp\{-\tau_i\cdot T^3\}\begin{pmatrix}\mathbf{n}_i\\ \mathbf{b}_i\\ \mathbf{t}_i\end{pmatrix}. \tag{2}$$

Here, T^2 and T^3 are two of the standard generators of three dimensional rotations, explicitly in terms of the permutation tensor we have $(T^i)^{jk}=\epsilon^{ijk}$. From Eqs. (1) and (2) we can compute the bond angles κ_i and the torsion angles τ_i using PDB data for $\mathbf{r}_i$. Alternatively, if we know these angles we can compute the coordinates $\mathbf{r}_i$. The common convention is to select the range of these angles so that κ_i is positive. In the continuum limit where Eq. (2) becomes the standard Frenet equation for a continuous curve, $\kappa_i\to\kappa(x)$ then corresponds to local curvature which is defined to be non-negative.

As a concrete example we now describe the 35 residue villin headpiece protein with PDB code 1VII that has been widely investigated, both theoretically and experimentally [4]. For example in the state-of-art simulation [5] succeeded in producing its fold for a short time with a root mean square distance (RMSD) accuracy of $\sim 2-3$ Å.

From the PDB data we compute the values of bond angles κ_i and torsion angles τ_i and the result is displayed in Fig. 1(a). when we use the (standard) convention that the discrete Frenet curvature κ is positive. In 1VII there are three α helices that are separated by two loops. When we use the PDB (NMR) convention for indexing the residues the first, longer, loop is located at sites 49–54 and the second, shorter, between 59–62.

We shall now show that Fig. 1(a) describes two soliton configurations, albeit in an encrypted form. In order to decrypt the data in Fig. 1(a) so that these solitons become unveiled we observe that the Eq. (2) has the following local $\mathbb{Z}_2$ gauge symmetry: At every site we can send

*chernodub@lmpt.univ-tours.fr
†shuangwei.hu@lmpt.univ-tours.fr
‡antti.niemi@physics.uu.se

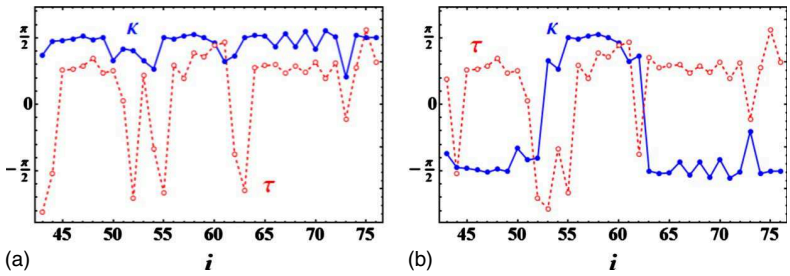

FIG. 1. (Color online) (a) (left): The bond and torsion angles of 1VII, computed with the standard convention that the discrete Frenet curvature κ is positive. (b) (right): The Z_2 gauge transformed bond and torsion angles.

$$\mathbb{Z}_2: \begin{cases} \kappa_i \to \kappa_i \cdot \cos(\Delta_i) \\ \tau_i \to \tau_i + \Delta_{i-1} - \Delta_i \end{cases} \tag{3}$$

and when we choose at each site $\Delta_i = 0$ or $\Delta_i = \pi$ where $\Delta_i = \pi$ is the nontrivial element of the $\mathbb{Z}_2$ gauge group, the Cartesian coordinates $\mathbf{r}_i$ computed from the discrete Frenet equation remain intact.

The gauge transformation that we introduce is a remnant of the continuum convention to choose the curvature $\kappa(x)$ to be always non-negative. As a consequence it can only be defined in a piecewise manner, between the straight segments and the conjugacy points where the curvature vanishes and the Frenet frame cannot be introduced. For a continuum curve, it can then become an issue how to determine $\kappa(x)$ through these conjugacy points so that its first derivative along the curve remains continuous. Instead of the common convention of a piecewise defined and non-negative curvature, which often leads to a discontinuous first derivative at the conjugacy points, we here use a definition where we allow $\kappa(x)$ to change its sign over points/segments where it vanishes, in such a manner that its first derivative along the curve remains continuous. This introduces a discrete gauge structure that ensures the equivalence of the two alternative descriptions.

For a discrete curve the continuity is not really an issue, if we define the bond angle κ_i to be non-negative the vanishing of κ_i does not pose a similar kind of a problem as in the continuum. But it turns out that by demanding κ_i to be non-negative, we translate the presence of a conjugacy point into sign changes in the torsion angle τ_i between adjacent sites. This allows us to easily locate the potential presence of a loop in the raw PDB data, since a (continuum) topological domain wall necessarily involves the presence of a conjugacy point.

If we implement the $\mathbb{Z}_2$ gauge transformation in the data displayed in Fig. 1(a), at the points where τ_i changes its sign between adjacent points, we arrive at the apparently quite different Fig. 1(b). Unlike in Fig. 1(a), the profile of κ_i in Fig. 1(b) now clearly displays the hallmark profile of a topological soliton-(anti)soliton pair in a double-well potential: The two soliton profiles are located around the sites with indices 49–54 and 59–62 which are the locations of the two loops in 1VII. These profiles interpolate between the two "ground-state" values $\kappa_i \approx \pm \pi/2$ that pinpoint the locations of the α helices in 1VII. Moreover, the two downswings in the value of τ_i from the value $\tau_i \approx 1$ that mark the locations of the α helices, coincide with the locations of the two soliton profiles. The ensuing combined profile of κ_i and τ_i is qualitatively consistent with a double-well potential structure in the (κ, τ) plane that has the form displayed in Fig. 2: When we move from left to right in Fig. 1(b), we follow a trajectory in the (κ, τ) plane that starts by fluctuating around the potential energy minimum at $(\kappa, \tau) \approx (-\pi/2, 1)$ in Fig. 2, corresponding to the first α helix. The trajectory then moves through the first loop a.k.a. soliton (red line) to the second potential energy minimum i.e., α helix at $(\kappa, \tau) \approx (+\pi/2, 1)$ in Fig. 2, and finally back through the second loop a.k.a. soliton (blue line) to the first potential energy minimum at $(\kappa, \tau) = (-\pi/2, 1)$.

We now describe a theoretical model introduced in [9,10] that reproduces the (κ, τ) profile in Fig. 1(b) as a combination of two soliton solutions to its equations of motion, with a very high accuracy for the central α carbons. The model is defined by the energy functional

$$E = \sum_{i=1}^{N-1} (\kappa_{i+1} - \kappa_i)^2 + \sum_{i=1}^{N} c \cdot (\kappa_i^2 - m^2)^2 + \sum_{i=1}^{N} \{ b\kappa_i^2 \tau_i^2 + d\tau_i + e\tau_i^2 + q\kappa_i^2 \tau_i \}. \tag{4}$$

Here N is the number of central α carbons and (c, m, b, d, e, q) are parameters. We refer to [9,10] for a detailed motivation of Eq. (4): The first sum describes nearest neighbor interactions along the protein. The second sum describes a local self-interaction of the bond angles. The third sum describes local interactions between bond and torsion

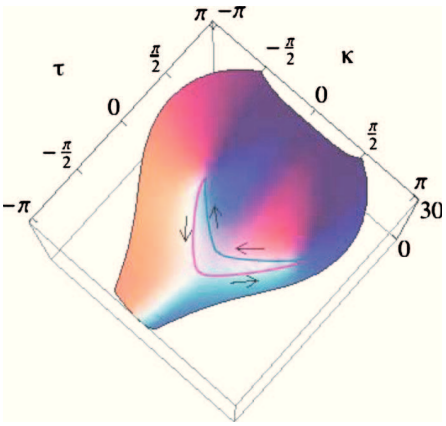

FIG. 2. (Color) The potential energy on (κ, τ) plane that corresponds qualitatively to the data in Fig. 1(b), the soliton between sites 49–54 corresponds to the red trajectory and the soliton between sites 59–62 to the blue trajectory.

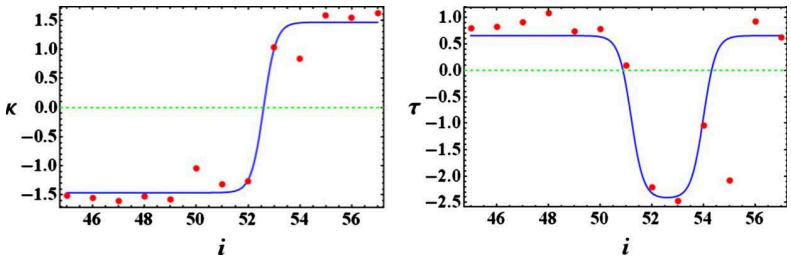

FIG. 3. (Color online) The PDB data for the first α-helix-loop-α-helix motif in 1VII, on the left κ_i and on the right τ_i, together with the least square approximations Eqs. (5) and (6) (solid blue line).

angles, its first term has an origin in a Higgs effect which is due to the potential term in the second sum. The second term in the third sum is the Chern-Simons term, it is responsible for the chirality of the protein chain. The third term is a Proca mass term and the last term can also be related to the Abelian Higgs model, and it is also chiral. As explained in [10] this energy functional is essentially unique, and in particular it can be related to a gauge invariant (supercurrent) version of the energy of 1+1 dimensional lattice Abelian Higgs model. In three space dimensions this model is also known as the Ginzburg-Landau model of conventional superconductivity [2].

We fully appreciate that the detailed fold of a given protein is determined by the specifics of its unique amino acid sequence. The interactions that contribute to the fold include hydrophobic, hydrophilic, long-range Coulomb, van der Waals, saturating hydrogen bonds and so forth interactions [11]. Consequently, *a priori* a given protein should not be approximated by a homopolymer model.

Note that in Eq. (4) there is no reference to the specifics of the interactions that are presumed to drive the folding process. The only explicit long-range force present in Eq. (4) is the nearest-neighbor interaction described by the first term. Moreover, as it stands Eq. (4) depends only on *six* site-independent, homogeneous parameters. There is no direct reference whatsoever to the underlying in general highly inhomogeneous amino acid structure of a protein. We argue that this becomes possible since Eq. (4) supports *solitons* that describe the common secondary structural motifs such as α-helix/β-strand-loop-α-helix/β-strand as solutions to its classical equations of motion. Furthermore, even though the actual numerical values of the parameters are certainly motif dependent and for long loops that constitute bound states of several solitons one might need to introduce more than six parameters, we expect there to be wide *universality* so that a given soliton with its relatively few parameters describes a general class of homologous motifs. Consequently only a relatively small set of parameters is needed to provide soliton templates for structure prediction. In fact, we propose that solitons are the mathematical manifestation of the experimental observation, that the number of different protein folds is surprisingly limited. The presence of solitons could then be the reason for the success of bioinformatics based homology modeling in predicting native folds [4].

In order to quantitatively disclose the soliton solution of Eq. (4) we start by observing that the first two sums in Eq. (4) can be interpreted as a discrete version of the energy of the 1+1 dimensional double well $\lambda\phi^4$ model that is known to support the topological kink soliton. In the continuum limit the kink soliton has the analytic form [1,2],

$$\kappa(x) = m \cdot \tanh(m\sqrt{c} \cdot [x - x_0]).$$

We can try to estimate the parameters m and c for each of the two solitons in the Fig. 1(b) by a least square fitting where we use this continuum soliton to approximate the exact soliton solution of the discrete equations of motion. We consider here explicitly only the first soliton of 1VII, located between (PDB index) sites 49–54. We assume that the discretized kink-soliton describes the profile of κ_i, and using the sites 46–56 we find the following least square fit

$$\kappa(x) \approx 1.4627 \cdot \tanh(2.0816[x - 52.597]). \tag{5}$$

In order to construct $\tau(x)$ we solve for its equation of motion in Eq. (4). Up to the parameters the dependence of τ_i on the kink soliton is then uniquely determined by the model, and the result is

$$\tau(x) \approx -2.4068 \cdot \frac{1 - 0.4689 \cdot \kappa^2(x)}{1 - 0.4619 \cdot \kappa^2(x)}. \tag{6}$$

In Fig. 3 we show how the data in Fig. 1(b) is described by the approximate soliton profile Eqs. (5) and (6). When we construct the ensuing discrete curve in the three dimensional ambient space by solving Eq. (2) with for κ_i and τ_i given by Eqs. (5) and (6), we reproduce the first loop of 1VII with a surprisingly good RMSD accuracy of $\sim$1.4 Å for the PDB indices 46–56. We think that this is quite remarkable, in particular by taking into account the simplicity of our approximation: Our Ansatz depends on *only one single function*, the hyperbolic tangent, that is determined by Eq. (4). In addition, there are the six parameters in Eq. (4). But a minimum of six characteristic parameters are needed to describe any loop configuration, and each of these can be given a very definite interpretation. The parameters are as follows:

(1) The location of the soliton along the protein (in κ_i)
(2) The size of the soliton in number of sites
(3) The asymptotic value of κ_i away from the soliton
(4) The asymptotic value of τ_i away from the soliton
(5) The value of τ_i at the center of the soliton
(6) The relative position of κ_i and τ_i for the center of soliton

For both (3) and (4) there are two possible values, corresponding to α helix and β strand. For (6) we have found that the location of the center of the soliton is slightly different in the variables κ_i and τ_i.

We take the remarkable success of our construction Eqs. (5) and (6) to be a strong argument in support of universality in protein folding. The same set of six parameters should describe corresponding loops in *any* homologically related

protein. Obviously this needs to be confirmed, and we are now in the process of constructing the explicit soliton profiles for several homologically related proteins in the PDB.

In order to construct a more accurate description of 1VII, we resort to a numerical construction of a soliton solution to the equations of motion if Eq. (4). We use simulated annealing that involves a Monte Carlo energy minimization of the energy functional

$$F = -\beta_1 \cdot \sum_{i=1}^{N} \left\{ \left(\frac{\partial E}{\partial \kappa_i} \right)^2 + \left(\frac{\partial E}{\partial \tau_i} \right)^2 \right\} - \beta_2 \cdot \sqrt{\frac{1}{N} \sum_{i=1}^{N} |\mathbf{r}_{PDB}(i) - \mathbf{r}_{soliton}(i)|^2}. \qquad (7)$$

with a simultaneous cooling of the two (inverse) temperatures β_1 and β_2. Here, the first sum vanishes when we have a solution to the classical difference equation of motion of Eq. (4), the cooling simulates a gradient flow toward the critical points i.e., classical solutions of Eq. (4). Since Eq. (4) can have several different critical points, we introduce the second term that computes the RMSD distance between the ith α carbon of the solution and the protein we wish to construct. The second term in Eq. (7) then acts like a chemical potential that selects the parameters in Eq. (4) so that we arrive at a soliton solution that corresponds to the actual, given protein fold.

We have numerically constructed the classical solutions of Eq. (4) that describe the secondary structural motifs in proteins with PDB codes 1VII, 2RB8 and 3EBX. The first one has three α helices separated by loops, while the second and third have β-strand-loop-β-strand motifs. Both cases can be described equally by Eq. (4), the only difference is that in the case of β strands the two minima of the (classical) potential in Eq. (4) are located at $(\kappa, \tau) \approx (\pm 1, \pi)$. In each of the proteins that we have studied we have routinely been able to reproduce the secondary structural motifs as classical soliton solutions to the equations of motion for Eq. (4) in terms of only six parameters and with an overall RMSD accuracy of less than 1.0 Å per motif which is essentially the experimental accuracy in x-ray crystallography and NMR; in our simulations the first sum in Eq. (7) decreases typically by around ten orders of magnitude indicating that the final configuration is a solution, essentially within numerical accuracy. Consequently at least in these proteins the secondary structural motifs can be viewed as solitons of the model Eq. (4), within experimental accuracy. Since the motifs that we have considered are quite generic in PDB data, we have very little doubt that our results will continue to persist whenever we have loops that connect α helices and/or β strands. And as long as the loops are not very long and do not describe bound states of several solitons there does not appear to be any need to introduce more than six parameters. Work is now in progress to systematically construct and classify the solitons that describe the secondary structural motifs in a large class of biologically active proteins.

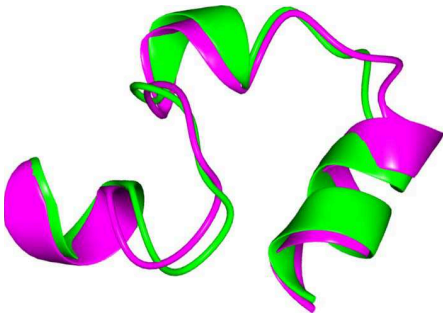

FIG. 4. (Color online) The helix-loop-helix-loop-helix structure of the 1VII protein (light grey, green online) together with its reconstruction in terms of two solitons (dark grey, purple online). The RMSD distance between the two configurations is ≈1.2 Å.

We have also made tentative attempts to use our solitons to reconstruct entire proteins, by *naively* joining the solitons that describe the secondary structural motifs at their ends. In the case of 1VII we have been able to reproduce in this manner the entire protein as a classical soliton with an overall RMSD accuracy of around 1.2 Å and the result is shown in Fig. 4. Even though the accuracy we obtain is very good, the loss of accuracy from ~0.7 to ~1.2 Å when we combine the two solitons in this particular case, suggests that we can still substantially improve the method of assembling an entire folded protein from its solitons. Work is now in progress to develop more efficient methods for assembling entire proteins from their solitons.

In conclusion, we have proposed that the common secondary structural motifs that describe loops connecting α helices and/or β strands can be interpreted as topological solitons, with the α helices and β sheets viewed as ground states that are interpolated by the loops as solitons. Entire proteins can then be assembled simply by combining these solitons together one after another. We have also presented a model that allows us to describe folded proteins in terms of its solitons within experimental accuracy. In its simplest form that we have considered here, the model describes a loop in terms of a single function and six site independent but in general motif dependent parameters, each of which have a direct relation to the overall geometric characteristics of the loop. This observation that all the details and complexities of amino acids and their interactions can be summarized in so simple terms suggests the existence of wide universality in protein folding. It can be viewed as a mathematically precise formulation of the experimental observation that the number of protein conformations is far more limited than the number of different amino acid combinations. Finally, we leave it as a future challenge to expand the model so that it incorporates an order parameter that describes the local orientation of the amino acids along the α carbon backbone.

Our research was supported by grant from the Swedish Research Council (V.R.). We thank Martin Lundgren for discussions.

[1] T. Dauxois and M. Peyrard, *Physics of Solitons* (Cambridge University Press, Cambridge, England, 2006).
[2] N. Manton and P. Sutcliffe, *Topological Solitons* (Cambridge University Press, Cambridge, England, 2004).
[3] A. S. Davydov, J. Theor. Biol. **38**, 559 (1973).
[4] K. A. Dill, O. S. Banu, M. S. Shell, and T. R. Weikl, Ann. Rev. Biophys. **37**, 289 (2008).
[5] G. Jayachandran, V. Vishal, and V. S. Pane, J. Chem. Phys. **124**, 164902 (2006).
[6] K. Huang, *Lectures On Statistical Physics And Protein Folding* (World Scientific, Singapore, 2005).
[7] H. M. Berman, K. Henrick, H. Nakamura, and J. L. Markley, Nucleic Acids Res. **35**, D301 (2007).
[8] P. J. Flory, *Statistical Mechanics of Chain Molecules* (Wiley, New York, 1969).
[9] A. J. Niemi, Phys. Rev. D **67**, 106004 (2003).
[10] U. H. Danielsson, M. Lundgren, and A. J. Niemi, e-print arXiv:0902.2920.
[11] See for example K. Cahill, Phys. Rev. E **72**, 062901 (2005).

Chapitre 8

Discrete nonlinear Schrodinger equation and polygonal solitons with applications to collapsed proteins

Nora Molkenthin, Shuangwei Hu, and Antti J. Niemi

Physical Review Letters 106, 078102 (2011)

Discrete Nonlinear Schrödinger Equation and Polygonal Solitons with Applications to Collapsed Proteins

Nora Molkenthin,[1,2] Shuangwei Hu,[1,2] and Antti J. Niemi[1,2]

[1]*Laboratoire de Mathematiques et Physique Theorique CNRS UMR 6083, Fédération Denis Poisson, Université de Tours, Parc de Grandmont, F37200, Tours, France*

[2]*Department of Physics and Astronomy, Uppsala University, P.O. Box 803, S-75108, Uppsala, Sweden*

(Received 7 October 2010; published 16 February 2011)

We introduce a novel generalization of the discrete nonlinear Schrödinger equation. It supports solitons that we utilize to model chiral polymers in the collapsed phase and, in particular, proteins in their native state. As an example we consider the villin headpiece HP35, an archetypal protein for testing both experimental and theoretical approaches to protein folding. We use its backbone as a template to explicitly construct a two-soliton configuration. Each of the two solitons describe well over 7.000 supersecondary structures of folded proteins in the Protein Data Bank with sub-angstrom accuracy suggesting that these solitons are common in nature.

DOI: 10.1103/PhysRevLett.106.078102 PACS numbers: 87.15.Cc, 05.45.Yv, 36.20.Ey

The discrete nonlinear Schrödinger equation [1] is a prime example of a universal equation. It originally appeared in the connection of polarons in molecular crystals [2] but has since had numerous applications from fiber optics and nonlinear acoustics to quantum condensates and ocean waves. The equation supports both stationary and time dependent solitons that were first introduced to describe Davydov solitons in proteins [3], then found in applications to the crystalline state of acetanilide [4], and subsequently emerged in Bose-Einstein condensates [5]. Today the discrete nonlinear Schrödinger equation together with its generalizations (GDNLS) form a very actively studied family of nonlinear equations, widely utilized to describe a multitude of phenomena in disparate physical, chemical, and biological scenarios [1–6].

In this Letter we argue that solitons of GDNLS equation are also common in polymers, they may even be pivotal in describing the collapsed phase: In general, a polymer such as protein displays three different *nontrivial* phases. These are in the universality class of self-avoiding random walk, in the universality class of Brownian motion, and in the universality class of a collapsed polymer [7]. The first two phases are theoretically quite well understood, and several models have been presented to describe them [8]. But the collapsed phase is much more difficult to describe and tractable models are hard to come by. Here we introduce a novel GDNLS equation that relates to an energy function that has been shown to characterize the collapsed phase [9,10]. We propose that the presence of solitons is essential for describing collapsed (chiral) polymers. While the model we consider is applicable for a large class of (chiral) polymers as a concrete example we address the problem of proteins in their native state, in particular, since there is a large amount of data available for comparisons [11].

We describe a polymer by the coordinates $\mathbf{r}_i$ of the N backbone carbons ($i = 1, \ldots, N$), in the case of proteins these coordinates can be downloaded from the Protein Data Bank (PDB) [12]. We compute the tangent vectors

$$\mathbf{t}_i = \frac{\mathbf{r}_{i+1} - \mathbf{r}_i}{|\mathbf{r}_{i+1} - \mathbf{r}_i|}. \tag{1}$$

The binormal and normal vectors are given by

$$\mathbf{b}_i = \frac{\mathbf{t}_{i-1} \times \mathbf{t}_i}{|\mathbf{t}_{i-1} - \mathbf{t}_i|} \quad \text{and} \quad \mathbf{n}_i = \mathbf{b}_i \times \mathbf{t}_i.$$

These vectors are subject to the discrete Frenet equation

$$\begin{pmatrix} \mathbf{n}_{i+1} \\ \mathbf{b}_{i+1} \\ \mathbf{t}_{i+1} \end{pmatrix} = \exp\{-\kappa_i T^2\} \exp\{-\tau_i T^3\} \begin{pmatrix} \mathbf{n}_i \\ \mathbf{b}_i \\ \mathbf{t}_i \end{pmatrix} \tag{2}$$

where T^2 and T^3 are two of the standard generators of three-dimensional rotations, explicitly in terms of the permutation tensor we have $(T^i)_{jk} = \epsilon^i_{jk}$.

From (1) and (2) we compute the bond angles κ_i and the torsion angles τ_i in terms of the PDB data for $\mathbf{r}_i$. Alternatively, if κ_i and τ_i are given we can compute the coordinates $\mathbf{r}_i$. The common convention is to select κ_i to be non-negative, the zeros of its continuum version (the curvature) correspond to the inflection points of the ensuing curve.

We determine κ_i and τ_i by locating the critical points of the following energy function [9,10],

$$E = -\sum_{i=1}^{N-1} 2\kappa_{i+1}\kappa_i + \sum_{i=1}^{N} \{2\kappa_i^2 + c(\kappa_i^2 - m^2)^2\} + \sum_{i=1}^{N} \{b\kappa_i^2\tau_i^2 + d\tau_i + e\tau_i^2 + q\kappa_i^2\tau_i\}. \tag{3}$$

We select κ_i to be periodic, $\kappa_i \in [-\pi, \pi] \bmod (2\pi)$. It is subject to both local and nearest-neighbor interactions. The variable $\tau_i \in [-\pi, \pi] \bmod (2\pi)$ is only subject to local

interactions. Finally, (b, c, d, e, m, q) are *global* parameters that in applications to folded proteins are specific to a given supersecondary structure, but are quite independent of the detailed monomer structure.

The energy function (3) is a discretized version of the standard Abelian Higgs Model; see [9] for details. The third term is a symmetry breaking potential. The closely related second term is a remnant of the method we have used to discretize second order derivatives and the fourth term has its origin in the familiar Higgs effect. The fifth term is a one-dimensional version of the Chern-Simons functional; its presence provides a very simple explanation of homochirality with a positive (negative) parameter e giving rise to right-handed (left-handed) chirality. The sixth term is a Proca mass, and the last term is a regulator; if this term is removed, the energy function (3) is exactly the Hamiltonian of a discrete Abelian Higgs Model with Chern-Simons term and Proca mass, in supercurrent variables that are commonly introduced in applications to superconductivity [9].

We note that if we delete all but the first term in the second sum, we arrive at the (discrete) Kratky-Porod model [13] of semiflexible polymers. It cannot describe the collapsed phase of polymers and, in particular, it does not support solitons.

In [10] it has been proposed that the critical points of (3) yield solitons, and approximative methods were introduced to describe them as models of supersecondary helix-loop-helix structures. We now show that (3) relates directly to the GDNLS equation. This equation emerges as follows: We first eliminate the auxiliary variable by varying the energy functional with respect to τ_i. This gives us an equation of motion to resolve for τ_i in terms of κ_i,

$$\frac{\partial E}{\partial \tau_i} = 2b\kappa_i^2\tau_i + 2e\tau_i + d + q\kappa_i^2 = 0 \Rightarrow \tau_i[\kappa_i] = -\frac{1}{2}\frac{d + q\kappa_i^2}{e + b\kappa_i^2}. \tag{4}$$

We then perform a variation of the energy functional with respect to κ_i, and substitute $\tau_i[\kappa_i]$ from (4) into the ensuing equation of motion to arrive at our GDNLS equation

$$\kappa_{i+1} - 2\kappa_i + \kappa_{i-1} = U'[\kappa_i]\kappa_i \equiv \frac{dU[\kappa]}{d\kappa_i^2}\kappa_i \quad (i = 1, \ldots, N) \tag{5}$$

(with $\kappa_0 = \kappa_{N+1} = 0$). This equation determines the stationary points of the following GDNLS Hamiltonian

$$H = -2\sum_{i=1}^{N-1}\kappa_{i+1}\kappa_i + \sum_{i=1}^{N}\{2\kappa_i^2 + U[\kappa_i]\}$$

where

$$U[\kappa] = -\left(\frac{bd - eq}{2b}\right)^2\frac{1}{e + b\kappa^2} - \left(\frac{q^2 + 8bcm^2}{4b}\right)\kappa^2 + c\kappa^4.$$

Here the second and the third term are familiar in the context of the nonlinear Schrödinger equation [1–6]. If only the third term is present the Hamiltonian relates to the Hasimoto representation of space curves [14]. Finally, the first term is a generalization of the Vinetskii-Kukhtarev potential [15] of nonlinear waveguides. But none of these truncations, even when they describe solitons, yield a model that relates to proteins in their native state.

If we choose the parameters in (3) so that the potential $U[\kappa]$ has two separate local minima, the results in [16] ensure the existence of a dark soliton solution that interpolates between these two minima. Such a qualitative form of $U[\kappa]$ typically follows if away from the vicinity of $\kappa = 0$ the potential becomes dominated by the second contribution to E in (3). This is the familiar double-well potential term, with minima at $\kappa = \pm m$. A dark soliton is then a configuration that interpolates from the ground state in the vicinity of $\kappa_1 \approx \pm m$ to the ground state in the vicinity of $\kappa_N \approx \mp m$ as we traverse the backbone. When we compute κ_i from (5) and τ_i from (4) and integrate the ensuing discrete Frenet equation we obtain a N-vertex polygonal chain such that a ground state with $\kappa \approx \pm m$ and τ given by (4) is a helix, with the dark soliton describing a loop that connects two helices.

We follow [16] to solve (5) iteratively by locating a fixed point of

$$\kappa_i^{(n+1)} = \kappa_i^{(n)} - \epsilon\{\kappa_i^{(n)}U'[\kappa_i^{(n)}] - (\kappa_{i+1}^{(n)} - 2\kappa_i^{(n)} + \kappa_{i-1}^{(n)})\}. \tag{6}$$

Here $\{\kappa_i^{(n)}\}_{i\in N}$ denotes the nth iteration of an initial configuration $\{\kappa_i^{(0)}\}_{i\in N}$ and ϵ is some sufficiently small but otherwise arbitrary numerical constant, for example, we can choose $\epsilon = 0.01$. It is obvious that a fixed point of (6) satisfies the GDNLS equation (5). As an initial configuration we utilize a step function, chosen to have the same overall topology as the desired dark multisoliton solution. Notice that as it stands, the energy functional (3) has the $\kappa \leftrightarrow -\kappa$ reflection symmetry that may not be exactly realized in applications to folded proteins, for example, there are proteins where a loop connects an α helix with a β sheet. Thus we explicitly break this symmetry using the parameter m: We set $m \to m_a$ for $N_{a-1} \le i \le N_a$ along the chain. Typical values for m_a are $m_a \approx \pm\pi/2$ for the α helix, and $m_a \approx \pm 1$ for the β strand.

We have performed extensive numerical investigations of the dark soliton solutions to (6). We have found that for proper values of the parameters solitons indeed exist and can be combined into multisolitons that together with (4) give a *very* high accuracy approximation of various folded protein structures that are stored in the PDB [12].

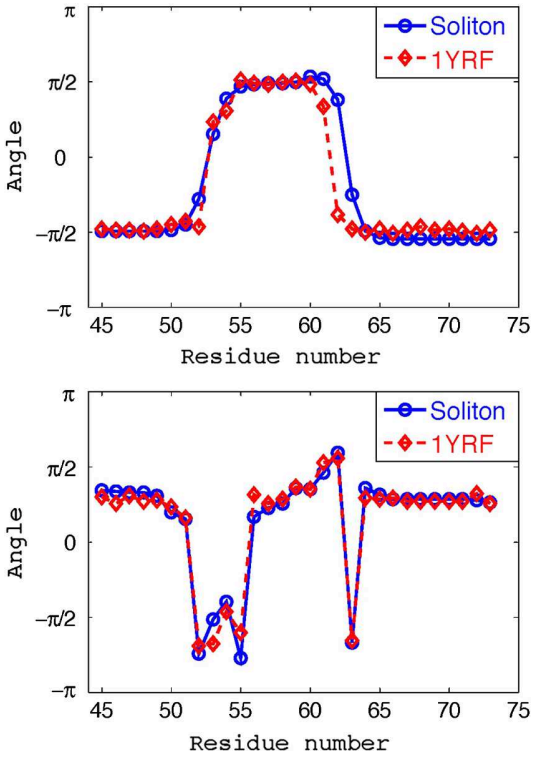

FIG. 1 (color online). (Top): The bond angles κ_i of 1YFR (red) for the sites 3–33 (45–78 in the PDB indexing convention) and their approximation by a soliton solution to Eq. (5) (blue). (Bottom): The torsion angles τ_i of 1YRF (red) for the sites 3–33 (45–78 in the PDB indexing convention) and their approximation by a soliton solution to Eq. (4) (blue).

As an example we construct two dark solitons using as our template the chicken villin headpiece subdomain HP35 (PDB code 1YRF) which is a naturally existing 35-residue protein. It has three α helices separated from each other by two loops. Together with the engineered version (2F4K in PDB) and the very similar HP36 (1VII in PDB), the HP35 has become the subject of very extensive studies both experimentally [17–20] and theoretically [21–24]. Using classical molecular dynamics, the authors of [21–24] report on the construction of native and near-native folds. The native fold in [23] deviates in average around 1.63 Å in C_α RMSD from the x-ray data [19] for the sites 2–34 (counting from the N terminus), and Ref. [24] reports very similar results with a proposed native fold average C_α RMSD around 1.54–1.65 Å for the sites 2–34. The overall resolution in the experimental x-ray data is 1.07 Å in RMSD [19]. We have selected this protein with the hope that by constructing it as a two-soliton solution with loops identified as the solitons, we can provide a new and beneficial perspective for molecular dynamics simulations to become even more effective.

In order to construct a two-soliton solution that describes the HP35 fold in PDB, we first convert the PDB coordinates for the C_α carbons to the bond and torsion angles using (2).

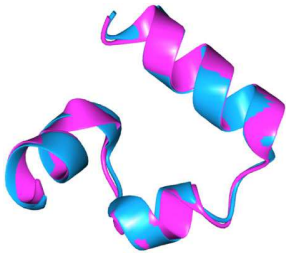

FIG. 2 (color online). Comparison between 1YRF backbone (red [dark gray]) and a soliton solution of (3) (blue [light gray]). The RMSD distance is 0.74 Å.

The result is shown in Fig. 1. The reason we do not consider the entire chain is that in order to compute these angles from the three-dimensional space coordinates we need to know the coordinates of three adjacent C_α carbons. From the κ_i profile we conclude that the C_α backbone of 1YRF consists of two dark solitons. These correspond to the two loops of 1YRF and are located around the sites 49–53 (PDB indexing) and 58–62 in Fig. 1, respectively. These solitons interpolate between ground states that correspond to the three α helices of 1YRF. The first helix is located between the sites 42–49, the second between the loops around sites 53–58, and the third occupies the remaining sites starting from 62 in Fig. 1. While the two-soliton profiles $\{\kappa_i\}$ are clearly identifiable, the profile of $\{\tau_i\}$ is substantially less regular and *a priori* one may expect that the strong irregularity in $\{\tau_i\}$ reflects the amino acid differences in the side chains. *Quite unexpectedly* we have found that this is not the case. The $\{\tau_i\}$ profile can be computed *very* accurately from (4) in terms of the soliton profile κ_i, as the apparent irregularity reflects *solely* the mod(2π) multivalued character of a periodic variable.

To construct the soliton profile for the entire chain, we introduce for each of the two would-be solitons the global parameters $(b, c, d, e, m_1, m_2, q)$: There is one set of parameters for the sites $i = 3$–13 (counting from N terminus) and another set of parameters for the remaining sites. We construct the ensuing soliton solution of (5) by iterating (6) to a fixed point, and compute its RMSD to 1YRF. We then change the parameters randomly and compute the new soliton profile, always starting from the same initial profile for the κ_i. We compare its RMSD to 1YRF with that obtained for the first set of initial parameters using the standard Metropolis algorithm devised to minimize RMSD. By repeating these steps in combination with simulated annealing we eventually produce our final soliton solution.

Note that even though we have seven parameters for each soliton, four of these are determined by the curvature and torsion on each side of the loop and thus *only three* parameters are needed for each of the loops.

Figure 2 compares our minimal RMSD two-soliton configuration with the 1YRF backbone constructed from the x-ray data, for the sites $i = 3$–33. The RMSD between the two configurations is 0.72 Å, well below the overall

TABLE I. Parameter values for a two-soliton solution that describe the entire 1YRF protein with accuracy 0.72 Å. We also present parameters for soliton-1 that describes the first loop (sites 2–13) with accuracy 0.75 Å, and t corresponding values for soliton-2 that describes the second loop (sites 14–33) with accuracy 0.28 Å.

parameter	b	c	d	e	q	m_1	m_2
1st set	$-3.070816340e-04$	$4.461893869e-01$	$1.142581922e-02$	$7.675000601e-04$	$-3.704049149e-03$	1.423206983	1.616099122
2nd set	$-1.095208557e-04$	1.172495797	$5.811514400e-04$	$2.013501270e-04$	$-2.880826898e-04$	1.520126333	1.540139296
soliton-1	$1.800314201e-04$	0.4222887366	$7.02765265e-03$	$4.663610215e-04$	$-2.190120515e-03$	1.444455611	1.565166201
soliton-2	$-2.22159366e-04$	1.088046084	$1.308858509e-03$	$3.94423507e-04$	$-6.4844084e-04$	1.518466566	1.543914339

resolution of the experimental x-ray data (which is 1.07 Å). Indeed, our dark two-soliton pair describes the native 1YRF backbone with an accuracy comparable to that of the radius of a carbon atom. In Table I we provide the parameter values for this configuration. We also present the parameter values for the best individual solitons that we have independently constructed for the two loops.

Since the solitons we have constructed employ the specific profile of 1YRF as a template, one might think that the parameter values in Table I are specific to this particular protein, reflecting its unique amino acid structure. However, this is *not* the case. For example, for the second soliton in Table I we find that there are presently a total of 7.736 unique supersecondary structures in the PDB with RMSD deviation less than 1.0 Å.

In conclusion, we have presented a novel generalized discrete nonlinear Schrödinger equation that supports solitons that describe chiral polymers such as proteins in their collapsed phase. The equation involves only *global* parameters, in particular, the fold is determined by a *single* function. With the 1YRF backbone as a template, we have constructed a soliton configuration that describes the backbone with an atomary level accuracy less than the radius of a carbon atom. Furthermore, we have found that *thousands* of supersecondary structures in the PDB are described with sub-angstrom accuracy by our solitons. Among the future challenges is the enumeration and modeling of the different supersecondary structures in the PDB and developing a relation between genome and a soliton basis of the PDB data.

Our research is supported by a grant from the Swedish Research Council (VR). We thank F. Sha and A.Khorotkin for help in comparing our solitons with the PDB data, M. Lundgren for discussions, and B. Malomed for pointing out a relation to the Vinetskii-Kukhtarev equation. N. M. thanks M. Herrmann for communications.

[1] P. G. Kevrekidis, *The Discrete Nonlinear Schrödinger Equation: Mathematical Analysis, Numerical Computations and Physical Perspectives* (Springer-Verlag, Berlin, 2009).

[2] T. Holstein, Ann. Phys. (N.Y.) **8**, 325 (1959).

[3] A. C. Scott, Phys. Rep. **217**, 1 (1992).

[4] J. C. Eilbeck, P. S. Lomdahl, and A. C. Scott, Phys. Rev. B **30**, 4703 (1984).

[5] J. C. Eilbeck and M. Johansson, *The Discrete Nonlinear Schrödinger Equation-20 years on, in Localization and Energy Transfer in Nonlinear Systems*, edited by L. Vázquez, R. S. MacKay, and M. Paz Zorzano (World Scientific, Singapore, 2003).

[6] A. C. Scott, *Nonlinear Science: Emergence and Dynamics of Coherent Structures* (Oxford University Press, Oxford, 2003), 2nd ed..

[7] P. G. De Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, 1979).

[8] J. F. Marko and E. D. Siggia, Phys. Rev. E **52**, 2912 (1995).

[9] U. H. Danielsson, M. Lundgren, and A. J. Niemi, Phys. Rev. E **82**, 021910 (2010).

[10] M. Chernodub, S. Hu, and A. J. Niemi, Phys. Rev. E **82**, 011916 (2010).

[11] K. A. Dill, O. S. Banu, M. S. Shell, and T. R. Weikl, Annu. Rev. Biophys. **37**, 289 (2008).

[12] H. M. Berman, K. Henrick, H. Nakamura, and J. L. Markley, Nucleic Acids Res. **35**, D301 (2007).

[13] O. Kratky and G. Porod, J. Colloid Sci. **4**, 35 (1949).

[14] H. Hasimoto, J. Fluid Mech. **51**, 477 (2006).

[15] V. O. Vinetskii and N. V. Kukhtarev, Sov. Phys. Solid State **16**, 2414 (1975).

[16] M. Herrmann, Applicable Analysis **89**, 1591 (2010).

[17] C. J. McKnight, P. T. Matsudaira, and P. S. Kim, Nat. Struct. Biol. **4**, 180 (1997).

[18] J. Meng, D. Vardar, Y. Wang, H. C. Guo, J. F. Head, and C. J. McKnight, Biochemistry **44**, 11 963 (2005).

[19] T. K. Chiu, J. Kubelka, R. Herbst-Irmer, W. A. Eaton, J. Hofrichter, and D. R. Davies, Proc. Natl. Acad. Sci. U.S.A. **102**, 7517 (2005).

[20] L. Wickstrom, Y. Bi, V. Hornak, D. P Raleigh, and C. Simmerling, Biochemistry **46**, 3624 (2007).

[21] G. Jayachandran, V. Vishal, and V. S. Pande, J. Chem. Phys. **124**, 164902 (2006).

[22] D. L. Ensign, P. M. Kasson, and V. S. Pande, J. Mol. Biol. **374**, 806 (2007).

[23] H. Lei and Y. Duan, J. Mol. Biol. **370**, 196 (2007).

[24] P. L. Freddolino and K. Schulten, Biophys. J. **97**, 2338 (2009).

Chapitre 9

Towards quantitative classification of folded proteins in terms of elementary functions

Shuangwei Hu, Andrei Krokhotin, Antti J. Niemi, and Xubiao Peng

Physical Review E 83, 041907 (2011)

Towards quantitative classification of folded proteins in terms of elementary functions

Shuangwei Hu,[1,2,*] Andrei Krokhotin,[2,†] Antti J. Niemi,[1,2,‡] and Xubiao Peng[2,§]

[1]*Laboratoire de Mathematiques et Physique Theorique CNRS UMR 6083, Fédération Denis Poisson, Université de Tours, Parc de Grandmont, F-37200, Tours, France*

[2]*Department of Physics and Astronomy, Uppsala University, P.O. Box 803, S-75108, Uppsala, Sweden*

(Received 26 November 2010; published 11 April 2011)

A comparative classification scheme provides a good basis for several approaches to understand proteins, including prediction of relations between their structure and biological function. But it remains a challenge to combine a classification scheme that describes a protein starting from its well-organized secondary structures and often involves direct human involvement, with an atomary-level physics-based approach where a protein is fundamentally nothing more than an ensemble of mutually interacting carbon, hydrogen, oxygen, and nitrogen atoms. In order to bridge these two complementary approaches to proteins, conceptually novel tools need to be introduced. Here we explain how an approach toward geometric characterization of entire folded proteins can be based on a single explicit elementary function that is familiar from nonlinear physical systems where it is known as the kink soliton. Our approach enables the conversion of hierarchical structural information into a quantitative form that allows for a folded protein to be characterized in terms of a small number of global parameters that are in principle computable from atomary-level considerations. As an example we describe in detail how the native fold of the myoglobin 1M6C emerges from a combination of kink solitons with a very high atomary-level accuracy. We also verify that our approach describes longer loops and loops connecting α helices with β strands, with the same overall accuracy.

DOI: 10.1103/PhysRevE.83.041907 PACS number(s): 87.15.A−, 87.15.Cc

I. INTRODUCTION

Comparative protein classification schemes such as CATH [1] and SCOP [2] are among the most valuable and widely employed tools in bioinformatics-based approaches to protein structure. These schemes classify folded proteins in terms of their geometric shape, starting from prevalent secondary structures such as α helices and β strands. But at the moment the final stages of the classification usually involve manual curation, and consequently these schemes are best suited for qualitative analysis of folded proteins.

The goal of the present article is to start development of novel tools that we propose can eventually provide a firm quantitative basis for the existing protein classification schemes. Ultimately we hope to close gaps between bioinformatics-based protein structure classification and physics-based atomary-level approaches to protein folding, to comprehensively address a wide range of issues such as protein structure prediction and relations between shape, function, and dynamics. In this way we hope to open doors to new ways of performing evolutionary, energetic, and modeling studies.

Our approach is based on the recent observation [3,4] that the geometric shape of helix-loop-helix motifs can be captured by a single elementary function that is familiar from the physics of nonlinear systems, where it describes the kink soliton. This function involves only a relatively small set of global parameters but still characterizes an entire supersecondary structure involving two (α) helices and/or (β) strands in addition to the loop that connects them. In Ref. [3] only individual supersecondary structures in relatively simple proteins and with quite short loops were considered. The approach proposed there did not work very well for entire protein chains, involving several helices and loops; it was essentially limited to a relatively short single loop with adjoining helices. The purpose of the present article is to show that the method can be developed to describe an *entire* protein and not just its helix-loop-helix segments. The protein can also be quite complex: it can involve several loops, both short and long and including those that connect α helices with β strands. Furthermore, the original ansatz can be even simplified without affecting its accuracy. Remarkably we observe no loss of accuracy even when the length and complexity of the protein chain increases. Indeed, there does not appear to be any limitations whatsoever that have to be imposed on the complexity of the protein for our approach to remain practical.

Our motivation derives from an investigation of nonlinearities that are generic in the force fields employed in classical molecular dynamics, a technique that is widely used in various theoretical studies of the structure, dynamics, and thermodynamical properties of proteins, and in determining their folding patterns in x-ray crystallography and NMR experiments [5]. A classical molecular dynamics approach such as AMBER [6] and GROMACS [7] describes the evolution of a folding protein in terms of Newton's law that determines the time dependence of the atomary spatial coordinates $\mathbf{X}(t) = \{\mathbf{x}_i(t)\}$:

$$m_i \ddot{\mathbf{x}}_i(t) \;=\; -\nabla_i U(\mathbf{X}). \tag{1}$$

Here $i = 1, \ldots, N$ catalog the individual atoms both in the protein molecule and in its environment, and $U(\mathbf{X})$ is an empirically constructed potential energy that governs the relevant mutual interactions between all atoms involved.

*shuangwei.hu@lmpt.univ-tours.fr
†andrei.krokhotine@cern.ch
‡antti.niemi@physics.uu.se
§xubiaopeng@gmail.com

Generically the potential energy is written as the sum of two terms [6]:

$$U(\mathbf{X}) = \sum U_{\text{covalent}}(\mathbf{X}) + \sum U_{\text{rest}}(\mathbf{X}). \tag{2}$$

The first term describes the covalent two-, three-, and four-body interactions between all covalently bonded atoms. The second term describes the noncovalent interactions between all atoms. For example, in the widely used harmonic approximation the two-body contribution to potential energy that describes the vibrational motion of all pairs of covalently bonded atoms acquires the familiar form

$$U^{(2)}_{\text{bond}} = \sum_{\text{bonds}} k_{ij}(|\mathbf{x}_i - \mathbf{x}_j| - r_{0ij})^2, \tag{3}$$

where r_{0ij} are the equilibrium distances between the pairs of covalently bonded atoms i and j, and k_{ij} are the ensuing spring constants.

But there are also nonlinear corrections to the potential energy such as (3), albeit in practice they may be difficult to account for in a systematic manner. The study of these nonlinearities forms a basis of the present work.

We start with a *Gedanken* experiment where we scrutinize a highly simplified version of an improvement to the harmonic approximation (3), with only a single (relative) coordinate on a line x so that Newton's equation is merely

$$m\ddot{x} = -\frac{dV}{dx},$$

where the potential has the form

$$V(x) = \tfrac{1}{2}k(x)\cdot(x-a)^2 \approx \tfrac{1}{4}\kappa\,(x+b)^2\cdot(x-a)^2.$$

That is, we account for nonlinear deviations from the harmonic approximation by promoting the spring constant to an x-dependent quantity. The equilibrium position $x = a$ of the harmonic approximation is recovered when $|x| \approx |a| << |b|$,

$$V(x) \;\approx\; \frac{1}{4}\kappa\, b^2(x-a)^2\cdot\left[1+\mathcal{O}\left(\frac{x}{b}\right)\right],$$

but here we retain the full potential. We introduce

$$c = -\tfrac{1}{2}(b+a)$$

and define

$$y = x - \tfrac{1}{2}(a-b)$$

to arrive at the familiar "$\lambda\phi^4$" (kink) equation of motion

$$\ddot{y} = -\frac{\kappa}{m}y(y^2-c^2)$$

with the explicit dark soliton solution

$$\begin{aligned} y(t) &= c\cdot\tanh\left[c\sqrt{\frac{\kappa}{2m}}(t-t_0)\right] \\ \Rightarrow\; x(t) &= y(t) + \frac{1}{2}(a-b) \\ &= -\frac{b\cdot e^{c\sqrt{\frac{\kappa}{2m}}(t-t_0)} - a\cdot e^{-c\sqrt{\frac{\kappa}{2m}}(t-t_0)}}{\cosh\left[c\sqrt{\frac{\kappa}{2m}}(t-t_0)\right]}. \end{aligned} \tag{4}$$

This is the hallmark dark soliton (kink) configuration that interpolates between the two uniform ground states at $x = a$ and $x = -b$ when $t \to \pm\infty$. The parameters a, b, t_0, and the combination $c\sqrt{\frac{\kappa}{2m}}$ are the canonical ones that characterize the asymptotic values of $x(t)$, *i.e.,* minima of the potential and the size and location of the soliton. It is also noteworthy that for finite t the soliton (4) describes a configuration with an energy above the uniform ground state $x \equiv a$ (or $x \equiv b$) but that nevertheless can not decay into $x \equiv a$ (or $x \equiv b$) through any kind of continuous finite energy transformation: A soliton configuration such as (4) cannot be obtained from any approach that accounts only for perturbations that describe small localized fluctuations around the uniform background ground state.

We argue that our example is not just an academic exercise but can be developed into a systematic tool to quantitatively characterize the geometrical shape of supersecondary structures such as helix-loop-helix motifs. In fact, we propose that the *very same function* (4) with t a length parameter that measures distance along a static protein backbone, together with its asymmetric generalization of the form

$$\tilde{x}(t) = \frac{b\cdot e^{c_1(t-t_0)} - a\cdot e^{-c_2(t-t_0)}}{e^{c_1(t-t_0)} + e^{-c_2(t-t_0)}}, \tag{5}$$

which becomes handy, e.g., when we consider loops connecting an α helix with a β strand, can describe the geometry of native folds of proteins in Protein Data Bank (PDB) [8]. Besides the four canonical soliton parameters that we have specified, we need to introduce only two additional independent global parameters to characterize a given supersecondary structure such as a helix-loop-helix motif and even an entire folded protein, with an atomary-level accuracy that matches the resolution in experimental data.

As an explicit example we have chosen myoglobin, a widely studied oxygen-binding protein of both historical and biological interest that has been discussed extensively in most biochemistry texts. Specifically, we have selected the 153 amino acid myoglobin with Protein Data Bank code 1M6C whose all-atom structure is known to an all-atom resolution of 1.90 Å in root-mean-square distance (RMSD) from x-ray diffraction measurements [8]. We analyze it in detail to show that its entire fold can be encoded into the global parameters of the elementary function (4), (5) with a RMSD accuracy of 1.27 Å for the central C_α carbons. Moreover, as the myoglobin involves only supersecondary structures with α and 3/10 helices that are connected by relatively short loops, we also verify that our approach can be extended to longer loops and loops that connect α helices with β strands. For this we analyze an α helix-loop-β strand segment in the HIV-1 reverse transcriptase protein with PDB code 3DLK. The loop is now clearly longer than those in myoglobin, but nevertheless we find that it can be described with comparable RMSD accuracy by the profile (5).

II. MYOGLOBIN AS MULTISOLITON

In order to describe the PDB fold of a relatively complex protein such as the 153 amino acid 1M6C in terms of the *single* elementary function (4), we start by computing the values of its discrete Frenet curvature κ_i and Frenet torsion τ_i from the PDB data. The relevant equations are as follows (for detailed derivation, see Ref. [9]): From PDB we get the

three dimensional coordinates $\mathbf{r}_i$ of the central α carbons ($i = 1, \ldots, N$). With these we compute the tangent vector $\mathbf{t}_i$ and the binormal vector $\mathbf{b}_i$ using

$$\mathbf{t}_i = \frac{\mathbf{r}_{i+1} - \mathbf{r}_i}{|\mathbf{r}_{i+1} - \mathbf{r}_i|} \quad \text{and} \quad \mathbf{b}_i = \frac{\mathbf{t}_{i-1} \times \mathbf{t}_i}{|\mathbf{t}_{i-1} - \mathbf{t}_i|}, \tag{6}$$

and the normal vector is given as

$$\mathbf{n}_i = \mathbf{b}_i \times \mathbf{t}_i.$$

These three vectors are subject to the discrete Frenet equation

$$\begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \end{pmatrix}_{i+1} = \exp(-\kappa_i \cdot T^2) \cdot \exp(-\tau_i \cdot T^3) \begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \end{pmatrix}_i . \tag{7}$$

Here T^2 and T^3 are two of the standard adjoint generators of three-dimensional rotations; explicitly in terms of the permutation tensor we have

$$(T^i)^{jk} = \epsilon^{ijk}.$$

From (6) and (7) we can compute κ_i and τ_i as the bond angles and the torsion angles in terms of the PDB data for $\mathbf{r}_i$. Alternatively, if we know these angles we can compute the coordinates $\mathbf{r}_i$ up to global rotations and translations. The common convention is to select the range of these angles so that κ_i is nonnegative. In the continuum limit where (7) becomes the standard Frenet equation for a continuous curve, $\kappa_i \to \kappa(x)$ then corresponds to local curvature, which is by convention defined to be nonnegative.

For 1M6C we take i to take values $i = 3, \ldots, 149 = N$; We leave out three (four) sites at both ends as we need three sites to initiate the computation of the κ_i and τ_i along the polygon, and the end points are anyway presumed to be subject to relatively large conformational fluctuations. In Fig. 1 (top) we display the κ_i and τ_i along the myoglobin backbone, using the standard differential geometric convention that κ_i is nonnegative.

Figure 1 displays the geometric structure of the 1M6C backbone fold: At the location of the α and 3/10 helices both κ_i and τ_i have pretty constant values, as expected for helical geometry. The difference between these two types of helices is visible in the figure, in (slight) difference in the corresponding constant values of κ_i and τ_i. At the location of loops, we note small variations in κ_i while the values of τ_i are fluctuating quite wildly. In order to identify the locations of the inflection points that determine the center of the loops, *i.e.,* solitons, we follow Ref. [3] and subject the data in Fig. 1 (top) to local $\mathbb{Z}_2$ gauge transformations in the loop regions; these transformations leave the solution of (7) intact and thus have no effect on the geometry of the space polygon. The result is shown in Fig. 1 (bottom); the two data point sets in the top and bottom of Fig. 1 describe the same space polygon. But from the bottom of Fig. 1 we conclude that in terms of κ_i we may interpret the backbone as a space polygon with 11 helices that are separated by ten inflection points (soliton centers); these are the points where κ_i changes its sign. Consequently we divide the backbone into ten supersecondary structures, each consisting of a helix-loop-helix soliton motif. These motifs are identified in Table I.

We note that PDB lists 1M6C as an eight-helix protein. But Fig. 1 reveals that there is an advantage to interpreting it in terms of a curve with ten inflection points, so that for a match with the functional form (4) we need to introduce ten overlapping segments. Furthermore, an examination of the PDB data reveals that there are four different types of loops, *i.e.,* solitons: Those that connect two α helices, those that

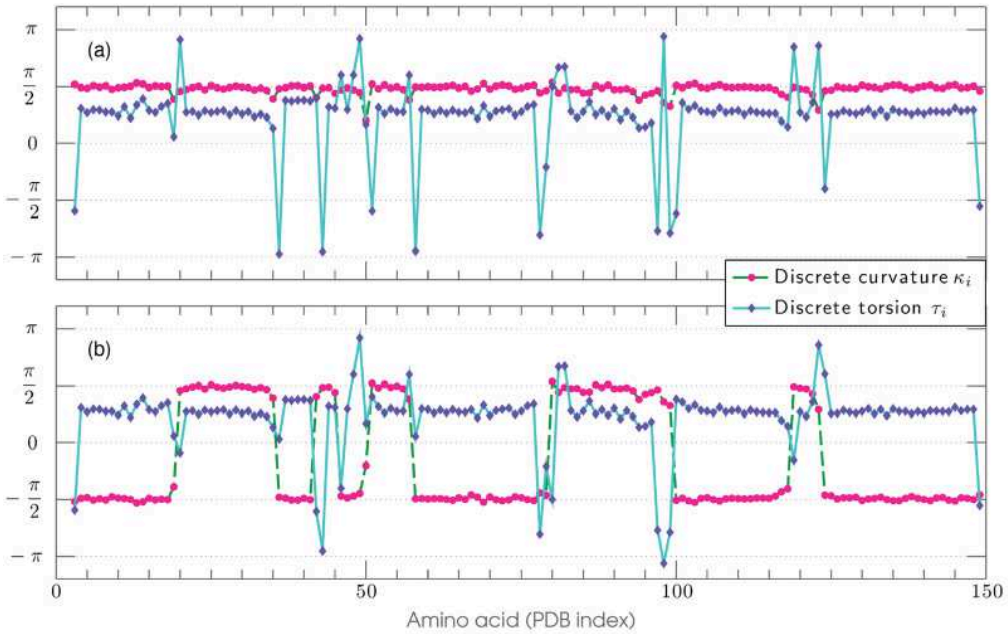

FIG. 1. (Color online) Values of κ_i and τ_i for 1M6C, obtained from PDB. (Top) These values using the standard convention that κ_i is nonnegative. (Bottom) The soliton structure resolved using $\mathbb{Z}_2$ gauge structure of the Frenet equation, by allowing κ_i to change sign whenever there is an inflection point. This identifies the soliton structures (loops) along the backbone. The indexing refers to the position of amino acids along the backbone, counting from the N terminus.

TABLE I. Parameters for solitons along the 1M6C C_α backbone, with indexing starting from the N terminus. The solitons have some overlap with their nearest neighbors. This enables us to combine them smoothly into a single multisoliton profile. The type identifies whether the soliton consists of a loop that connects α helices and/or 3/10 helices.

Soliton	1	2	3	4	5
Sites	3–24	22–42	37–46	43–50	47–58
Type	α-α	α-3/10	3/10-3/10	3/10-3/10	3/10-α
b_r	78.398	79.1807	68.7412	39.727	55.9241
c_r	1.5708	2.5280	2.5290	2.5550	3.1391
d_r	−0.2905	−0.1268	−0.2347	−0.2464	−0.2998
m_{r1}	1.53668	1.4979	1.56503	1.55474	1.5668
m_{r2}	–	1.5113	–	–	1.5651
s_r	20.5981	36.488	43.3982	45.657	51.733
RMSD	0.83	0.49	0.15	0.56	0.40
Soliton	6	7	8	9	10
Sites	52–80	59–98	81–119	102–123	120–150
Type	α - α	α-α	α-α	α-α	α-α
b_r	73.358	92.551	48.059	114.599	93.2733
c_r	2.1488	2.1874	1.95991	2.2796	2.5496
d_r	−0.3035	−0.4649	−0.3688	−0.1887	−0.1565
m_{r1}	1.52541	1.52732	1.48823	1.55946	1.54715
s_r	57.8112	80.7367	98.2245	118.8551	124.404
RMSD	1.12	1.46	1.62	0.60	0.37

connect an α helix with a 3/10-helix or vice versa, and finally those that connect two 3/10 helices.

In order to describe a motif consisting of a loop together with the two similar types of helices that it connects, we use the ansatz (4) with the symmetric ($a = b$) relation for the two parameters in (4). But for motifs where a loop connects two different types of helices (α with 3/10) we allow these parameters to be independent, reflecting the difference in the helices. Thus our ansatz for the entire backbone is the modification (5) of the ansatz introduced in Ref. [3]: For the bond angles we introduce the dark soliton profile

$$\kappa_i \;=\; (-1)^{r+1}\frac{m_{r1}\cdot e^{c_r(i-s_r)} - m_{r2}\cdot e^{-c_r(i-s_r)}}{2\,\cosh[c_r(i-s_r)]}, \qquad (8)$$

and we obtain the torsion angles from this soliton profile using the relation

$$\tau_i \;=\; -\frac{1}{2}\frac{b_r}{1+d_r\kappa_i^2}. \qquad (9)$$

Here $r = 1,\ldots,10$ labels the ten helix-loop-helix motifs of 1M6C, $(c_r, m_{r1}, m_{r2}, s_r)$ are the canonical parameters for a kink soliton, and (b_r, d_r) are additional parameters needed to express τ_i in terms of κ_i.

Note that (8) and (9) are *not* an *ad hoc* ansatz but can be firmly justified in terms of the equations of motion in an underlying Hamiltonian model that is based on the Abelian Higgs Model [10]. Indeed, at the level of the Abelian Higgs Model each of the parameters has a well-established interpretation in terms of charge, mass, self-coupling, *etc*. Here these parameters characterize the global attributes of a supersecondary helix-loop-helix structure: *A priori*, the two helices are described by two parameters each. These are the parameters m_{r1} and m_{r2} that determine the curvature of the two adjacent helices, and together with b_r, d_r we compute the torsions of these helices from (9). This leaves us with *only* two global parameters to determine the loop, in addition to the location parameter s_r of the inflection point. These two parameters characterize the overall curvature and torsion length of the loop, on both sides of the inflection point. Consequently, at the level of number of parameters our ansatz imposes that a loop involves no more parameter degrees of freedom than a helix. We emphasize that the ansatz involves only the *single* function (4), in its discrete form.

Since all the solitons except 2 and 5 connect similar helices, whenever $r \neq 2,5$ we can set $m_{r1} = m_{r2}$ while for solitons number 2 and 5 that connect two different kind of helices we choose $m_{r1} \neq m_{r2}$.

In our computations we determined the parameters using a standard Metropolis algorithm in combination with simulated annealing, to minimize the RMSD between the polygon described by our ansatz and the C_α backbone of the 1M6C protein in PDB. The actual algorithm is a very simple and straightforward application of standard Monte Carlo minimization that can be run even with a PC.

The construction of the 1M6C backbone proceeds in steps: We first construct the individual solitons. The ensuing residues and parameter values are given in Table I; note that each of the neighboring solitons has at least three common residues. This enables us to combine the solitons by smoothly continuing from one a set of values of (κ_i, τ_i) to the next one. Thus the entire backbone is built up from its elementary solitons, very much like children use interlocking plastic bricks such as Lego to build various objects. In Table I we display the parameters that yield the smallest RMSD value (RMSD = 1.27 Å) that we have obtained when we have subjected the entire 1M6C backbone to a RMSD minimization.

We also give the lowest RMSD values that we obtain for each of the individual solitons. For the solitons 1,2,3,4,5,9, and 10 we find very low RMSD values, clearly smaller than the radius (∼0.7 Å) of an individual carbon atom. However, the number of sites that appear in the solitons 3,4,5 are also quite small. This is due to the proximity of the ensuing solitons along the backbone. For solitons number 6,7,8 the RMSD values are somewhat larger, but the solitons are also longer. However, even in these cases our RMSD values are clearly below the overall 1.90 Å resolution in the underlying PDB data. In Fig. 2 we display the C_α backbone of 1M6C, together with its reconstruction in terms of the ansatz (8) and (9).

III. LONG LOOPS

The previous interpretation and construction of the myoglobin 1M6C backbone clearly demonstrates that the method proposed in Ref. [3] can be extended from helix-loop-helix supersecondary structures to entire proteins, even for relatively long proteins and with several helix-loop-helix combinations and both α and 3/10 helices. However, the question remains whether the quality of the method becomes adversely affected

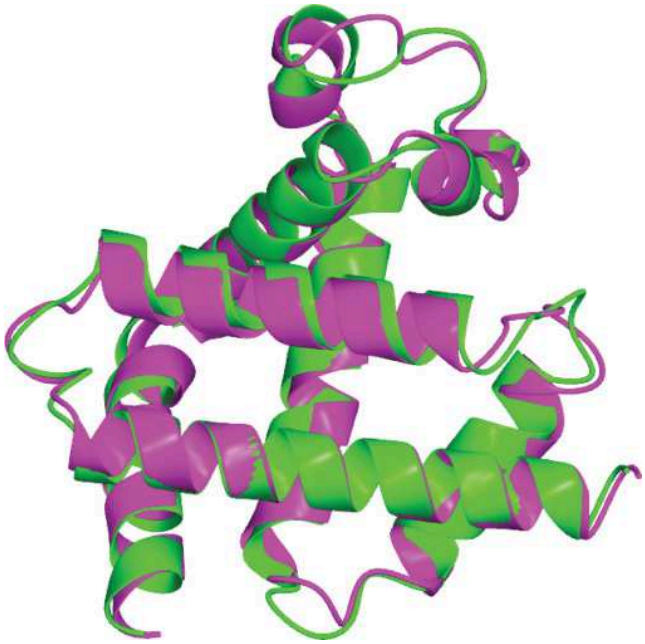

FIG. 2. (Color online) The structure of the 1M6C protein (green) together with its reconstruction in terms of our ansatz (purple). The RMSD distance between the two configurations is ≈1.27 Å.

if the loop length increases, and whether the method also describes loops that connect an α helix with a β strand. We address these issues by considering a protein loop with 12 C_α carbons connecting an α helix with a β strand. More specifically, we consider the sites 398–416 in the HIV-1 reverse transcriptase protein with PDB code 3DLK. In line with the construction of the solitons in the case of myoglobin, we describe the supersecondary structure with the following variant (5) of the ansatz (8) and (9):

$$\kappa_i = \frac{m_1\, e^{c_1(i-s_0)} - m_2\, e^{-c_2(i-s_0)}}{e^{c_1(i-s_0)} + e^{-c_2(i-s_0)}}, \tag{10}$$

and we again obtain the torsion angles from this soliton profile using the relation

$$\tau_i = -\frac{1}{2}\frac{b}{1+d\kappa_i^2}. \tag{11}$$

The asymmetric choice (m_1,c_1) versus (m_2,c_2) reflects the difference between the α helix and β strand, and we now start the indexing by choosing $i = 1$ for the site 398. With the choice of parameters in Table II we find that the ansatz describes the 3DLK segment with a RMSD accuracy of 1.13 Å; notice that due to the presence of exponentials, for high accuracy it is imperative to include sufficiently many decimal points in the parameters. In Fig. 3 we display the original 3DLK segment, together with its soliton approximation.

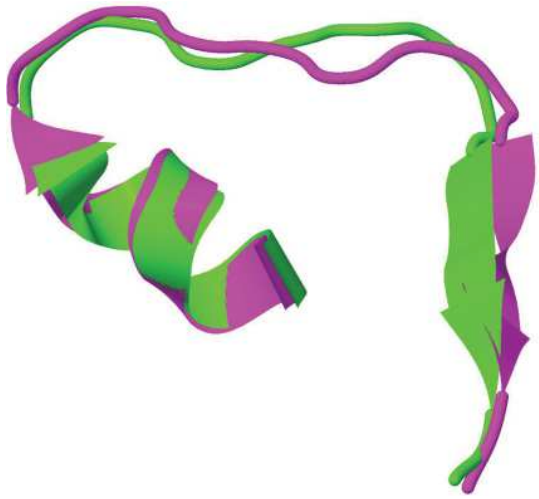

FIG. 3. (Color online) Sites 398–416 in 3DLK (green; PDB indexing) and their approximation (purple) by (10), (11) with parameter values given in Table II. The RMSD distance is ∼1.13 Å.

Thus the present approach is suitable not only for long protein chains such as myoglobin, but it also describes long loops and loops that connect very different kinds of helices such as α helices, 3/10 helices, and β strands. However, if the loop length increases substantially, we propose that a more accurate prescription is obtained by describing these loops as bound states of several short loops, each with the profile (10), (11). This is consistent with the well-known fact that short supersecondary structures are known to recur many times in PDB proteins. A detailed analysis of long loops as bound states of short loops (multisoliton states) will be presented elsewhere.

IV. CONCLUSION

Using the myoglobin 1M6C as an example, we have demonstrated that the entire native fold of a relatively long protein can be described with high accuracy as a combination of kink solitons, in a manner that involves only one single elementary function and only parameters that are *global* characteristics of the conformation. In this picture, each of the solitons describe a loop configuration that interpolates between two different helices. By inspecting a longer loop that connects an α helix with a β strand we have verified that the approach remains valid with no loss of accuracy as the loop size increases. However, for substantially longer loops, we expect that an interpretation in terms of a multisoliton configuration becomes more accurate both mathematically and phenomenologically.

The parameters that characterize a particular protein fold are all global, and specific to its supersecondary helix-loop-helix motifs. Consequently the determination of these

TABLE II. Parameters for describing the sites 398–416 along 3DLK. Indexing starts with $i = 1$ at site 398.

m_1	c_1	m_2	c_2	s_0	b	d
57.626008	1.836469	58.05348	1.8462217	10.43150	6 601 165.9	−0.000101

parameters becomes synonymous to a quantitative classification of proteins, and a general approach of parameter classification will be the subject of future publication. The presence of an underlying Hamiltonian interpretation also strongly suggests that our approach could eventually provide a bridge between comparative protein classification schemes such as CATH and SCOP, and physics-based approaches to protein folding and structure prediction, including folding pathways and various other dynamical issues that presently cannot be easily addressed in qualitative protein classification schemes. This should open doors to new ways of performing evolutionary, energetic, and modeling studies.

ACKNOWLEDGMENTS

We thank D. van der Spoel and R. Lavery for discussions and comments. This research has been supported by the Vetenskapsrdet Grant No. 2009-4099.

[1] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton, Structure **5**, 1093 (1997).

[2] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, J. Mol. Biol. **247**, 536 (1995).

[3] M. N. Chernodub, S. Hu, and A. J. Niemi, Phys. Rev. E **82**, 011916 (2010).

[4] N. Molkenthin, S. Hu, and A. J. Niemi, Phys. Rev. Lett. **106**, 078102 (2011).

[5] O. M. Becker, A. Mackerell, B. Roux, and M. Watanabe, *Computational Biochemistry and Biophysics* (Marcel Dekker, New York, 2001).

[6] J. W. Ponder and D. W. Case, Adv. Protein Chem. **66**, 27 (2003).

[7] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and J. Berendsen, Comput. Chem. **26**, 1701 (2005).

[8] H. M. Berman, K. Henrick, H. Nakamura, and J. L. Markley, Nucl. Acids Res. **35**, D301 (2007).

[9] S. Hu, M. Lundgren, and A. J. Niemi e-print arXiv:1102.5658 (2011).

[10] U. H. Danielsson, M. Lundgren, and A. J. Niemi, Phys. Rev. E **82**, 021910 (2010).

Conclusion

"Gardez les choses simples!" Un physicien n'oublie jamais le principe de simplicité, même face au kaléidoscope du monde des protéines. Dans cette thèse, nous avons présenté l'analyse des aspects *universels* des protéines repliées, au moyen de la géométrie différentielle, de la symétrie de jauge et de la théorie du soliton. Nos résultats les plus importants sont les suivants.

Nous avons montré que la validité d'un modèle à gros grain pour les protéines globulaires, basé sur la description de la géométrie différentielle discrète de la chaîne principale de la protéine, peut être étendue pour étudier sur l'existence d'une vaste universalité sur le niveau secondaire de la structure des protéines. La fonction d'énergie du modèle a été déterminée selon la ligne directive du principe d'invariance de jauge et a été utilisée dans des travaux précédents pour décrire l'universalité du niveau tertiaire, en termes de loi d'échelle du rayon de giration. La solution soliton du modèle, qui se manifeste sous la forme de motif hélice-boucle-hélice, a été maintenant montré pour saisir les caractéristiques globales de la structure super-secondaire. Les boucles, les configurations irrégulières dans la perspective conventionnelle, sont considérées au moment de jouer un rôle intrinsèque à combler les structures régulières des hélices α et des feuillets β.

L'application réussie du modèle soliton adoptée dans les présents travaux de thèse en ligne avec la découverte obtenue dans les modèles basés sur des méthodes de prédiction et sur des schémas de classification structurel qu'il y a un nombre limité des reploiements en état natif, dans leur nature modulaire simple. Cependant, comparé à la faiblesse de l'espace de recherche discrétisé dans les modèles basés sur des méthodes de prédiction, l'avantage du modèle de soliton actuel est de servir pour une méthodologie *fonctionnelle* de construction de la super-structure secondaire dans un espace continu, offrant ainsi la souplesse fondamentale pour analyser l'ensemble de la chaîne d'une protéine longue.

La description donnée par le modèle *simple* utilisé ici ne fournit pas de détails atomiques, en maintenant l'analyse à un niveau à grains grossiers. Pourtant, il s'agit d'une observation surprenante que l'orientation de la chaîne latérale, en termes d'atomes C_β, est entièrement déterminée par la configuration des longueurs, soulignant une fois de plus le rôle de la géométrie. Cela suggère l'utilité de la démarche actuelle de la modélisation plus détaillée des protéines, ainsi que dans la compréhension des principes physiques qui sous-tendent le problème du repliement des protéines.

Bien que notre cadre d'étude reste dans une représentation homopolymère de la structure protéique (la description *la plus simple*), l'accord excellent de notre modèle avec la protéine repliée suggère que le repliement des protéines est soumis à des contraintes géo-

métriques, du moins au stade précoce lorsque des interactions non covalentes ne sont pas encore dominées par la force motrice. L'introduction de nouvelles informations en acides aminés spécifiques, même dans un *simple* hydrophobe-hydrophile régime, doivent être compatibles avec la forme géométrique du squelette protéique. La concurrence entre eux seraient probablement utiles pour réussir à discriminer le pli cible correcte, ainsi que pour convenablement décrire la dynamique du processus de pliage. Ces considérations seraient les orientations des futurs travaux.

Par souci d'exhaustivité, il est prometteur de regarder l'image dynamique de repliement des protéines comme l'initialisation, la propagation et la terminaison des solitons, ainsi que la collision entre solitons multiple de toute la chaîne polypeptidique. Initialement, un ensemble d'acides aminés adjacents (un angle de liaison comprend trois acides aminés) prend soit hélice α soit le feuillet β comme etat fondamental due principalement à leur préférence de structure secondaire. Cet état fondamental va se propager le long de la chaîne, sans l'apport continu d'énergie, jusqu'à ce qu'elle se termine par un acide aminé polaire (appelée la coiffe de l'hélices). Le rôle de la coiffe de l'hélices est similaire avec le phenomene d'accrochage en theorie du soliton. La stabilité topologique et le comportement collectif des solitons assurent l'intégralité des éléments de structure secondaire, tandis que l'arrangement optimal de ces solitons multiples pour toute la chaîne est bien sûr tirée par l'interaction hydrophobe. Nous espérons que les travaux futurs rendront cette image dynamique claire.

"Keep things simple!" A physicist never forgets the simplicity principle, even facing with the kaleidoscope of protein world. In this thesis, we have presented the analysis of the *universal* aspects of folded proteins, by means of differential geometry, gauge symmetry and soliton theory. Our most important results are as follows.

We has shown that the validity of a coarse-grained model for globular proteins, based on discrete differential geometry description of the C_α-backbone, can be extended to investigate the existence of wide universality on the secondary level of protein structure. The energy function of the model is determined under the guideline of gauge invariance principle and had beed used in previous work to describe the universality on the tertiary level, in terms of scaling law of the gyration radius. The kink soliton solution of the model, manifested in the form of helix-loop-helix motif, has been now shown to capture the global characteristics of the super-secondary structure. Loops, the irregular configurations in conventional perspective, are regarded at the moment to play an intrinsic role in bridging the regular structures of α-helices and β-strands.

The successful application of the soliton model adopted in the present thesis works in line with the discovery obtained in template-based prediction methods and structural classification schemes that there is a limited number of native state folds, in their simple modular nature. However, compared with the weakness of discretized search space in template-based prediction methods, the advantage of present soliton model is to serve for a *functional* methodology of building the super-secondary structure in a continuous space, thus providing fundamental flexibility to analyze the entire chain of a long protein.

The description given by the *simple* model used here does not provide atomic details, keeping the analysis at a coarse-grained level. Yet it comes a surprising observation that the orientation of side-chain, in terms of C_β atoms, is entirely determined by the backbone configuration, emphasizing once more the role of geometry. It suggests the usefulness of present approach in the more detailed modeling of protein, as well as in the understanding of physical principles that underlie the protein folding problem.

Though our the framework remains within a homopolymer representation of the protein structure (the *simplest* description), the excellent fitting of our model with the folded protein suggests that protein folding is subject to the geometric constraints, at least at the early stage when the non-covalent interactions have not yet dominated the driving force. The further introduction of amino acid specific information, even within a *simple* hydrophobic-hydrophilic scheme, should be compatible with the shape geometry of the protein backbone. The competition between them would probably help to successfully discriminate the correct target fold, as well as to suitably describe the dynamics of folding process. These considerations would be the directions of the future work.

For the sake of completeness, it is promising to view the dynamical picture of protein folding as the initialization, propagation and termination of soliton, as well as the collision between multiple solitons along the whole polypeptide chain. Initially a set of adjacent amino acids (one bond angle involves three amino acids) takes either α-helix or β-sheet as ground state mainly due to their secondary structure preference. This ground state will propagate along the chain, without continuous input of energy, until it terminates by a special polar amino acid (called a helix cap). The role of a helix cap is similar with the pinning phenomenon in soliton theory. The topological stability and collective behavior of

soliton ensures the integrality of secondary structure elements, while the optimal arrangement of these multiple solitons for the whole chain is of course driven by the hydrophobic interaction. We hope the future work will make this dynamical picture clearer.

Annexes

Annexe A

Simulation technics

Here some technical details of simulation are summarized. The main objective is to simulate Eq. (7) in Chapter 7. We first discuss the parameter learning by means of so-called double optimization. Then we outline the Monte Carlo techniques, including both theoretical basis and simulation details. At last we give the formula of calculating the similarity measurement, Root Mean Squared Deviation (RMSD).

A.1 Parameter learning

In Paper 2, we define an objective functional to parameterize the energy functional, given the native protein structure. There are six parameters to be fit, i.e. $\Theta \equiv (c, m, b, d, e, q)$ (see Eq. (4) in the paper). A tempting approach of parameter learning would be to solve the minimization problem

$$\min_{\Theta} F(\Theta) = \mathrm{RMSD}\left(\mathbf{R}_{PDB}, \arg\min_{\mathbf{R}} E(\mathbf{R}; \Theta)\right), \tag{A.1}$$

where target protein structure $\mathbf{R}_{PDB} = \{\mathbf{r}_{PDB}(i), i = 1, \ldots N\}$ and energy minimum structure $\mathbf{R}$ is computed from the optimal bond angles and torsion angles of the energy functional E. There are two lays of minimization procedures in this approach ; given a set of parameter we firstly have to minimize the energy functional. This approach had been test and shown to be fraught with difficulties. First, this objective functional presumes that the target protein structure is the unique minimum of the energy functional. But the objective functional is shown to have many local minima and a small move in the parameter space may trigger very unpredictable change for the optimal structure. Second, the two layers of minimization require high computation cost. Third, calculating the derivative of the objective functional is extremely difficult (especially since discrete Frenet equation is involved), and thus direction search methods like gradient descent cannot be applied here. As a result, random search methods such as Monte Carlo would be required.

To overcome these difficulties, we propose a new objective functional which intrinsically

utilize the Monte Carlo random search (also see Eq. (7) in Chapter 7)

$$\begin{aligned} F(\Theta, \mathbf{R}) &= -\beta_1 \sum_{i=1}^{N} \left\{ \left(\frac{\partial E}{\partial \kappa_i} \right)^2 + \left(\frac{\partial E}{\partial \tau_i} \right)^2 \right\} \\ &\quad -\beta_2 \mathrm{RMSD}\left(\mathbf{R}_{PDB}, \mathbf{R}\left(\{\kappa_i, \tau_i\} \right) \right). \end{aligned} \tag{A.2}$$

Here β_1, β_2 are two inverse temperatures. The second term has an essential difference from Eq. (A.1) that the configuration $\mathbf{R}(\{\kappa_i, \tau_i\})= \{\mathbf{r}_i, i = 1, \ldots, N\}$ doesn't necessarily correspond to the minimum of the energy functional, given the parameter Θ. In other words, we promote both $\mathbf{R}$ (or equivalently $\{(\kappa_i, \tau_i)\}$) and Θ to be the variables of the objective functional. Thanks to the first penalty term, the generated configuration will finally arrive at the minimum of the energy functional, at the same time as the parameters are optimized. So we would call this approach as *double optimization*. It is worthy to remark that both Eq. (A.1) and (A.2) don't depend on the explicit form of the energy functional.

In practice, simulating annealing is used to speed up the search of global minima. It is tricky to design a simultaneous cooling for tuning the two inverse temperatures, β_1 and β_2. There is no universal rule but just trial and error. At the same time, it is quite helpful to reset the trial configuration as the perturbation of target structure if the second term keeps going "uphill".

A.2 Markov chain Monte Carlo

In the core of Monte Carlo simulation is the sampling technique, i.e. the generation of random numbers obeying desired distribution. Markov chain Monte Carlo (MCMC) provides a quite general algorithm of sampling, no matter the probability distribution is simple or complex. The idea is based on building a Markov chain that takes the desired distribution as its equilibrium distribution. In this section, we first introduce Markov chains and then show its application on global minimization, with emphasis on the trial move of protein system.

A.2.1 Markov chain

For a finite state space Ω, a Markov chain is a sequence of random variables (X_t) with the Markovian property, namely that, for all t, all $x_0, \ldots, x_t, x \in \Omega$, we require

$$\Pr\left(X_{t+1} = x | X_0 = x_0, \ldots, X_t = x_t \right) = \Pr\left(X_{t+1} = x | X_t = x_t \right). \tag{A.3}$$

In the time-homogeneous transition case, we denote the transition matrix as

$$W_{mn} \equiv \Pr\left(X_{t+1} = n | X_t = m \right) = \Pr\left(X_1 = n | X_0 = m \right). \tag{A.4}$$

Its stationary distribution π is in fact the eigenvector with eigenvalue 1 (all other eigenvalues should be less than 1),

$$\forall n \in \Omega,\ \pi_n = \sum_{m \in \Omega} \pi_m W_{mn}. \tag{A.5}$$

A Markov chain is called ergodic if any state can reach at other states after finite steps, i.e. there exists integer $s \geq 1$ such that for any $m, n \in \Omega$, $W^s_{mn} = \sum_{k\in\Omega} W_{mk} W^{s-1}_{kn} > 0$. The essential property of ergodic Markov chains ensure the existence and uniqueness of a stationary distribution, regardless of their initial state.

Theorem For a finite ergodic Markov chain, there exists a unique stationary distribution π such that for any $m, n \in \Omega$, $\lim_{s\to\infty} W^s_{mn} = \pi_n$.

In practice, if W satisfies the so-called detailed balance

$$\forall m, n \in \Omega,\ \pi_m W_{mn} = \pi_n W_{nm}, \tag{A.6}$$

such a π is a stationary distribution with respect to the transition matrix W. To prove, just sum over each side,

$$\sum_{m\in\Omega} \pi_m W_{mn} = \sum_{m\in\Omega} \pi_n W_{nm} = \pi_n \sum_{m\in\Omega} W_{nm} = \pi_n. \tag{A.7}$$

This is just the stationary equation (A.5).

Detailed balance provides us a convenient way to design samplers by associating Markov chains with appropriate stationary distributions. One popular choice is to apply the Metropolis algorithm. Given a symmetric trial matrix, $\alpha_{mn} = \alpha_{nm}$, the Metropolis transition probability matrix reads

$$W_{mn} = \begin{cases} \alpha_{mn}, & \pi_n \geq \pi_m, m \neq n \\ \dfrac{\pi_n}{\pi_m}\alpha_{mn} \equiv \rho_{mn}\alpha_{mn}, & \pi_n < \pi_m, m \neq n \\ 1 - \sum_{m'\neq n} W_{m'n}, & m = n \end{cases}. \tag{A.8}$$

Such a algorithm is a kind of rejection method—propose and accept/reject : It first proposes a move from $X_m \to X_n$ with probability α_{mn}. This trial is accepted if the probability of X_n is larger than the probability of X_m. Otherwise accept the trial with a probability ρ_{mn}. The last relation guarantees the normalization condition on the transition probability matrix. It is straightforward to verify that Metropolis algorithm satisfies the detailed balance. An advantage here is that the relative probability density ρ_{mn} is relatively easy to calculate compared to the absolute probability π, since the computation of its normalization factor is often nontrivial.

A.2.2 Monte Carlo minimization

If we translate an energy functional $E(X)$ into some distribution, say canonical distribution in statistical physics

$$\pi(X) = \frac{e^{-\beta E(X)}}{Z(\beta)}, Z(\beta) = \int_\Omega e^{-\beta E(X)} dX, \tag{A.9}$$

then the ground state will dominate the equilibrium state, especially at high inverse temperature β. This idea has been widely used for the search of global minimum. We will now explain how, in practice, the trial move from state $X = m$ to state $X = n$ is accepted or

rejected. First propose a random trial (see next section for details) according to α_{mn}. If $E(X=n) \leq E(X=m)$, we accept the trial. If $E(X=n) > E(X=m)$, we calculate

$$\rho_{mn} = \frac{\pi(X=n)}{\pi(X=m)} = e^{-\beta(E(X=n)-E(X=m))}. \tag{A.10}$$

To decide, we generate a random number ξ in the interval $[0,1]$ in a uniform distribution. *The probability that* $\xi \leq \rho_{mn}$ *equals* ρ_{mn}. Therefore accept the trial if $\xi \leq \rho_{mn}$; otherwise reject. This rule ensures that the acceptance of the trial $i \to j$ is the probability ρ_{mn}.

In the simulated annealing method, the temperature is gradually decreased during the simulation process, i.e. the inverse temperature β is increased. When β is small at very beginning, the current state changes almost randomly. But as β goes to infinity, the state of the system undergoes a increasingly "downhill" on the landscape of the energy functional. The allowance for "uphill" moves potentially helps the simulation escape from local optima. In practice, we use the exponential cooling scheme, i.e. $\beta_t = \gamma\beta_{t-1} = \gamma^t\beta_0, \gamma \in (0,1)$. The simulation of double optimization in Eq. (A.2) involves two (inverse) temperatures. We tune them so that both equally contribute to the total accept ration of the trial, saying fifty percents in practice.

A.2.3 Monte Carlo trial move

Due to the complexity of protein-like system, it becomes nontrivial to design a symmetric Monte Carlo move, under the line of matrix α_{mn}. Many investigations have shown that both local and nonlocal moves are necessary for the effective overcome of the local minima in MC simulation. Typically there are three kinds of moves as below. Before we give the formula, here we remark that for a discrete curve of length N the index of bond angle runs from 2 to $N-1$ and the index of torsion angle runs from 3 to $N-1$.

A.2.3.1 Crankshaft move

Select randomly two C_α's $i, j, i < j$, such that $j - i \leq n_c$, with $n_c \ll N$ (in practical cases we randomly choose $n_c \in [2,6]$). Then, rotate C_α's $i+1, \ldots, j-1$ of an angle $\Delta\phi_c$ around the axis $\mathbf{r}_j - \mathbf{r}_i$ (Fig. A.1). The angle $\Delta\phi_c$ is chosen randomly with a uniform probability distribution in the interval $[-\Delta\phi_m/2, \Delta\phi_m/2]$. This move has to be carried out in coordinate representation.

Denote $\mathbf{r}_i = (a,b,c), \mathbf{r}_j - \mathbf{r}_i = (u,v,w)$. Then after a rotation of an angle $\Delta\phi_c$, any point $\mathbf{r}_k = (x,y,z)$ will change into $\mathbf{r}'_k = (x',y',z')$

$$\begin{aligned} x' &= \frac{1}{L^2}\left[a\left(v^2+w^2\right) + u\left(-bv-cw+ux+vy+wz\right)\right. \\ &\quad + \left((x-a)\left(v^2+w^2\right) + u\left(bv+cw-vy-wz\right)\right)\cos\Delta\phi_c \\ &\quad \left. + L\left(bw-cv-wy+vz\right)\sin\Delta\phi_c\right], \end{aligned} \tag{A.11}$$

$$y' = x'\left(a\to b, b\to c, c\to a, u\to v, v\to w, w\to u\right), \tag{A.12}$$

$$z' = x'\left(a\to c, b\to a, c\to b, u\to w, v\to u, w\to v\right), \tag{A.13}$$

where $L = \sqrt{u^2+v^2+w^2}$.

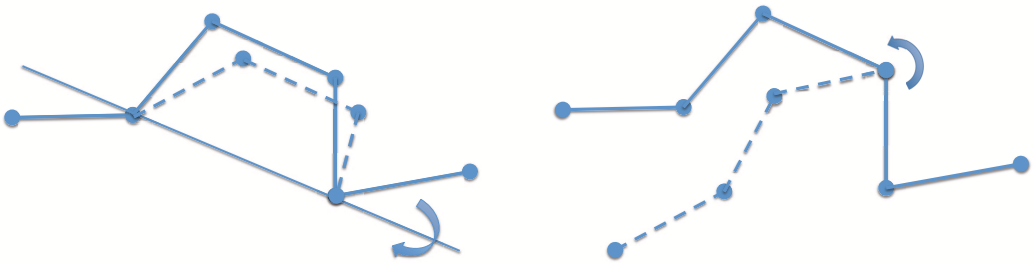

FIGURE A.1 – Monte Carlo move. Left : crankshaft move. Right : pivot move.

A.2.3.2 Reptation move

This kind of move cuts off a C_α from one end of the chain while appends a new C_α at the other end, whose orientation is chosen randomly. In the angle representation, in one case, to reshuffle angles, $\psi_i \leftarrow \psi_{i+1}$ for $2 \leq i \leq N-2$, and $\theta_i \leftarrow \theta_{i+1}$ for $3 \leq i \leq N-2$, and to assign a new value to ψ_{N-1} and θ_{N-1}. In the other case, $\psi_i \leftarrow \psi_{i-1}$ for $3 \leq i \leq N-1$, and $\theta_i \leftarrow \theta_{i-1}$ for $4 \leq i \leq N-1$, and to assign a new value to ψ_2 and θ_3. In both cases, new values for bond angles at ending C_α's are picked up randomly with a uniform probability distribution on the sphere, that is

$$\psi_{2,N-1} = \arccos \xi, \xi \sim \text{Unif}\,[0,1]\,. \tag{A.14}$$

A.2.3.3 Pivot move

Take randomly one C_α at the site i, with $1 \leq i \leq N-1$ as the pivot point, and then rotate the part of the chain coming after the pivot point while keeping invariant the other part of the chain (Fig. A.1). In the angle representation, this is simply implemented by updating either ψ_{i+1}, or θ_{i+1} (if $i = 1$ only ψ_2 is updated). Which angle is to be updated is again chosen randomly or both. The updating rule is $\theta_{i+1} \leftarrow \theta_{i+1} + \Delta\theta_p$, with the uniform distribution $\Delta\theta_p \sim \text{Unif}\,[-\Delta\theta_m/2, \Delta\theta_m/2]$. Similar rule applies for ψ_{i+1}.

A.2.3.4 Constraints

There are two main constraints for the move, the truncated region of bond angle and the self-avoiding condition. From the statistics of bond angles, we know they are restricted to the region $[30°, 100°]$ (see Fig. 2.5). Meanwhile, the self-avoiding condition reads

$$|\mathbf{r}_i - \mathbf{r}_j| \geq 3.7,\ \forall\ |i-j| \geq 2. \tag{A.15}$$

Both constraints origins from the steric effect and energy disfavor. If Monte Carlo move breaks either of these two constraints, the trial will be rejected. Modification can be tried in the opposite direction of the bad trial. Or directly modify the sampling ; for example, truncate the region of ξ in Eq. (A.14).

In general, the efficiency of Monte Carlo dynamics may depend crucially on tuning the different control parameters which we have introduced, $n_c, \Delta\phi_m, \Delta\theta_m, \Delta\psi_m$, possibly considering them as non-constant functions of the simulated annealing temperature. Efficiency also depends on the relative frequency of the different kinds of moves which we use. Again, we need trial and error to decide the optimal strategy.

A.3 Root Mean Squared Deviation

Given the coordinates of two proteins, $\mathbf{X} = \{\mathbf{x}_i, i = 1, \ldots, N\}$ and $\mathbf{Y} = \{\mathbf{y}_i, i = 1, \ldots, N\}$, we can define a quantity of simarity, Root Mean Squared Deviation (RMSD), as following

$$\mathrm{RMSD}\left(\mathbf{X}, \mathbf{Y}\right) = \min_{\mathbf{O},\mathbf{b}} \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left|\mathbf{O}\mathbf{x}_i + \mathbf{b} - \mathbf{y}_i\right|^2}. \tag{A.16}$$

This formula can be regarded as the split of the information for the coordinate pair into two separate parts, i.e. the intrinsic (shape) similarity RMSD and the freedom of the rotation matrix $\mathbf{O}$ and the translation vector $\mathbf{b}$. The explicit calculation of the minimization can be found via the method of singluar value decomposition (SVD) of matrix theory. If $\mathbf{X}$ and $\mathbf{Y}$ are both $3 \times N$ matrices and both mass centers of each structure are shifted to the origin ($\mathbf{b} = \mathbf{0}$), then we can calculate the 3×3 correlation matrix $\mathbf{C} = \mathbf{Y}\mathbf{X}^T$. Take the singular value decomposition, $\mathbf{C} = \mathbf{U}\mathbf{W}\mathbf{V}^T$, $\mathbf{U}$ and $\mathbf{V}$ are two orthogonal matrices, $\mathbf{W}$ the diagonal one. The action of $\mathbf{W}$ may be interpreted as some kind of stretching when applied to a vector. Since we are interested only in transformations which keep our structures rigid, we throw $\mathbf{W}$ away and retain only the two rotation matrices. This leaves us the desired optimal rotation $\mathbf{O} = \mathbf{U}\mathbf{V}^T$.

There is another way, called the quaternion method [37], to determine RMSD. We have chosen this method because it can be readily implemented, and faster than the above method. The formula is summarize as following. After both centroids are placed at the origin, calculate the correlation matrix $\mathbf{C} = \mathbf{Y}\mathbf{X}^T$. And then form a new 4×4 symmetric matrix

$$\begin{aligned}
\mathbf{D} = \left(\mathbf{D}_{ij}\right), \quad & \mathbf{D}_{ij} = \mathbf{D}_{ji}, \quad & \mathbf{D}_{11} = \mathbf{C}_{11} + \mathbf{C}_{22} + \mathbf{C}_{33}, \\
\mathbf{D}_{12} = \mathbf{C}_{23} - \mathbf{C}_{32}, \quad & \mathbf{D}_{13} = \mathbf{C}_{31} - \mathbf{C}_{13}, \quad & \mathbf{D}_{14} = \mathbf{C}_{12} - \mathbf{C}_{21}, \\
\mathbf{D}_{22} = \mathbf{C}_{11} - \mathbf{C}_{22} - \mathbf{C}_{33}, \quad & \mathbf{D}_{23} = \mathbf{C}_{12} + \mathbf{C}_{21}, \quad & \mathbf{D}_{24} = \mathbf{C}_{13} + \mathbf{C}_{31}, \\
\mathbf{D}_{33} = \mathbf{C}_{22} - \mathbf{C}_{11} - \mathbf{C}_{33}, \quad & \mathbf{D}_{34} = \mathbf{C}_{23} + \mathbf{C}_{32}, \quad & \mathbf{D}_{44} = \mathbf{C}_{33} - \mathbf{C}_{11} - \mathbf{C}_{22},
\end{aligned} \tag{A.17}$$

Then RMSD takes the value

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}\left(\mathbf{x}_i^2 + \mathbf{y}_i^2\right) - 2\lambda_{\max}}{N}}, \tag{A.18}$$

where $\lambda_{\max}$ is the largest eigenvalue of matrix $\mathbf{D}$. The corresponding rotation matrix $\mathbf{O}$ is given by

$$\mathbf{O} = \begin{pmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2\left(q_1 q_2 - q_0 q_3\right) & 2\left(q_1 q_3 + q_0 q_2\right) \\ 2\left(q_1 q_2 + q_0 q_3\right) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2\left(q_2 q_3 - q_0 q_1\right) \\ 2\left(q_1 q_3 - q_0 q_2\right) & 2\left(q_2 q_3 + q_0 q_1\right) & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{pmatrix}, \tag{A.19}$$

where $\mathbf{q} = (q_0, q_1, q_2, q_3)$ is the eigenvector of $\mathbf{D}$ with respect to the largest eigenvalue, $\mathbf{D}\mathbf{q}^T = \lambda_{\max}\mathbf{q}^T$. In practice, the largest eigenvalue of a matrix can be approximated by power iteration or Rayleigh quotient method.

We remark that given a curve $\mathbf{X} = \{\mathbf{r}_i, i = 1, \ldots, N\}$ and its perturbation $\mathbf{Y} = \{\mathbf{r}_i + \epsilon \cdot \mathbf{e}_i, i = 1, \ldots, N\}$ ($\mathbf{e}_i$ is random unit vector), the similarity measurement should be proportional to the perturbation amplitude ϵ. Besides RMSD, we could have other definitions of similarity, especially making use of the local Frenet frame. The so-called Frenet distance, explicitly independent of rotation and translation, is defined as

$$\rho(\mathbf{X}, \mathbf{Y}) = \sum_{i=3}^{N-1} \mathbf{u}_i^{(\mathbf{X})} \cdot \mathbf{u}_i^{(\mathbf{Y})}, \tag{A.20}$$

where the unit vector $\mathbf{u} = (\sin\psi\cos\theta, \sin\psi\sin\theta, \cos\psi)$, formed by the bond angle ψ and torsion angle θ.

Annexe B

Discrete Frenet equations

In this appendix we will show the possible ways of discretizing Frenet equations, with emphasizing the advantage of the choice in the papers. As an application, we give the explicit calculation on discrete $\alpha-$helix, as well as on $\beta-$sheet. Both share the same equation since $\beta-$sheet is just a deformed helix due to its intrinsic twist.

B.1 Discretization of Frenet equations

Firstly recall the continuous Frenet equations

$$\frac{d}{ds}\begin{pmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{pmatrix}=\begin{pmatrix}0 & \kappa & 0\\ -\kappa & 0 & \tau\\ 0 & -\tau & 0\end{pmatrix}\begin{pmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{pmatrix}\equiv\mathbf{Q}(s)\begin{pmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{pmatrix}. \tag{B.1}$$

The Frenet matrix $\mathbf{Q}(s)$ is skew-symmetric, $\mathbf{Q}^T=-\mathbf{Q}$. This differential system has the ordered exponential solution, with the form

$$\begin{pmatrix}\mathbf{t}(s)\\ \mathbf{n}(s)\\ \mathbf{b}(s)\end{pmatrix} = \mathbf{U}(s)\begin{pmatrix}\mathbf{t}(0)\\ \mathbf{n}(0)\\ \mathbf{b}(0)\end{pmatrix}, \tag{B.2}$$

$$\mathbf{U}(s) \equiv \mathcal{P}\left(\exp\left\{\int_0^s \mathbf{Q}(s')\,ds'\right\}\right) \tag{B.3}$$

$$= 1+\int_0^s ds'\mathbf{Q}(s')+\int_0^s ds'\int_0^{s'} ds''\mathbf{Q}(s')\,\mathbf{Q}(s'')+\cdots \tag{B.4}$$

It is important to notice that $\mathbf{U}(s)$ is an orthogonal transformation matrix, i.e. $\mathbf{U}^T=\mathbf{U}^{-1}$, which follows the exponentiation construction and $\mathbf{Q}^T=-\mathbf{Q}$. The orthogonality ensures that the Frenet basis are always orthogonal and normalized, i.e.

$$|\mathbf{t}(s)|^2+|\mathbf{n}(s)|^2+|\mathbf{b}(s)|^2=|\mathbf{t}(0)|^2+|\mathbf{n}(0)|^2+|\mathbf{b}(0)|^2. \tag{B.5}$$

The Taylor expansion of $\mathbf{U}(s)$ comes a closed form when $\kappa(s)$ and $\tau(s)$ are constant. The examples include the helices and the numerical case in which we discretize the arc length

with small step Δs so that $\kappa(s)$ and $\tau(s)$ become piecewise constant. In the latter case, we have the explicit expression within Δs

$$\mathbf{U} = e^{\mathbf{Q}\Delta s} \tag{B.6}$$

$$= \exp\begin{pmatrix} 0 & \kappa\Delta s & 0 \\ -\kappa\Delta s & 0 & \tau\Delta s \\ 0 & -\tau\Delta s & 0 \end{pmatrix} = \begin{pmatrix} a & b & c \\ -b & d & e \\ c & -e & f \end{pmatrix}, \tag{B.7}$$

$$a = \frac{\tau^2 + \kappa^2\cos(q\Delta s)}{q^2}, b = \frac{\kappa}{q}\sin(q\Delta s), c = (1-\cos(q\Delta s))\frac{\kappa\tau}{q^2}, \tag{B.8}$$

$$d = \cos(q\Delta s), e = \frac{\tau}{q}\sin(q\Delta s), f = \frac{\kappa^2 + \tau^2\cos(q\Delta s)}{q^2}, q = \kappa^2 + \tau^2. \tag{B.9}$$

This expression in fact lies on the basis of polyhelix model in protein research [26]. In this representation, C_α atoms are connected by helices, whose curvatures, torsions and arc length are nonlinearly fitted. The values of curvature and torsion can be used to characterize the structure preferences. There is, however, an simpler choice of representation, the impulse function for both curvature and torsion profile along the arc length, that is, having nonzero values only at vertices. This implies the polygonal representation of protein structure, as we have chosen. The advantage lies on the fact that arc length equals the distance between two consecutive C_α atoms, i.e. $s = i\delta, i = 0, \ldots, N$ (N is the number of C_α atoms). So the ordered exponential matrix only involve once $\mathbf{Q}$ to switch from one C_α atom to its neighbor,

$$\mathbf{U}((i+1)\delta) = e^{\mathbf{Q}(i\delta)}. \tag{B.10}$$

From the Trotter-Suzuki formula, one can show that

$$e^{\mathbf{Q}} = e^{\kappa\mathbf{T}_3 + \tau\mathbf{T}_1} = e^{\kappa\mathbf{T}_3}e^{\tau\mathbf{T}_1} + \mathcal{O}(\kappa\tau), \tag{B.11}$$

where $\mathbf{T}_1$ and $\mathbf{T}_3$ are generators of SO(3) rotation, $(\mathbf{T}_i)_{jk} = \epsilon^{ijk}$. It comes a surprise to see that the above approximation expression becomes exact when we choose a special discretization of Frenet frame [23], as following

$$\begin{cases} \mathbf{t}_i = \dfrac{\mathbf{r}_{i+1} - \mathbf{r}_i}{|\mathbf{r}_{i+1} - \mathbf{r}_i|}, \mathbf{b}_i = \dfrac{\mathbf{t}_{i-1} \times \mathbf{t}_i}{|\mathbf{t}_{i-1} \times \mathbf{t}_i|}, \mathbf{n}_i = \mathbf{b}_i \times \mathbf{t}_i, \\ \begin{pmatrix} \mathbf{t} \\ \mathbf{n} \\ \mathbf{b} \end{pmatrix}_i = e^{\kappa_i\mathbf{T}_3}e^{\tau_i\mathbf{T}_1}\begin{pmatrix} \mathbf{t} \\ \mathbf{n} \\ \mathbf{b} \end{pmatrix}_{i-1}. \end{cases} \tag{B.12}$$

This formula can be regarded as two-point difference of the continuous equations. For example, $\mathbf{t} = \mathbf{r}_s \to (\mathbf{r}_{i+1} - \mathbf{r}_i)/\delta$ and $\mathbf{b} = -\mathbf{n} \times \mathbf{t} \to (\mathbf{t}_{i-1} - \mathbf{t}_i) \times \mathbf{t}_i = \mathbf{t}_{i-1} \times \mathbf{t}_i$. The discretized curvature κ_i and torsion τ_i have now the geometrical meaning of bond angle and torsion angle ; see Chapter 6 for details. In Ref. [50, 51, 52], they proposed another approach which is essentially the three-point difference scheme, which makes the computation more complex and parameters less geometrical meaning. In Ref. [36], they used the discretization scheme by means of Cayley transform

$$\mathbf{U} \approx \left(1 + \frac{\delta}{2}\mathbf{Q}\right)\left(1 - \frac{\delta}{2}\mathbf{Q}\right)^{-1}. \tag{B.13}$$

It serves a good approximation if δ is small. When it comes to the application on proteins, such a scheme has no direct geometric meaning of curvature and torsion.

B.2 Discrete helix

In the publications, we pointed out that bond angle and torsion angle for an α-helix take the typical value $\pi/2$ and 0.879, respectively. Here we illustrate the calculation on discrete α-helix, with parameters be consistent with protein structure in PDB. Taking helix axis as z-axis, the protein chain is described by its position vector

$$\mathbf{r}_i = \left(R\cos\left(i\Delta\phi\right), R\sin\left(i\Delta\phi\right), i\Delta z\right), \tag{B.14}$$

where the angular step $\Delta\phi = 100^\circ$, the longitudinal step $\Delta z = 5.4\text{Å}/3.6 = 1.5\text{Å}$, and R is the helix radius. Statistics from PDB shows that the virtual bond length between two consecutive C_α atoms is almost constant (no matter in helices or not),

$$\delta^2 = \left|\mathbf{r}_{i+1} - \mathbf{r}_i\right|^2 = 3.8^2 = 2R^2\left(1 - \cos\Delta\phi\right) + \Delta z^2. \tag{B.15}$$

This condition decides the value of the helix radius,

$$R = \sqrt{\frac{\delta^2 - \Delta z^2}{2\left(1 - \cos\Delta\phi\right)}} = \sqrt{\frac{3.8^2 - 1.5^2}{2\left(1 - \cos 100^\circ\right)}} = 2.28\text{Å}. \tag{B.16}$$

Now we construct the discrete Frenet frame as

$$\begin{aligned}
\mathbf{t}_i &= \frac{\mathbf{r}_{i+1} - \mathbf{r}_i}{\delta} && \text{(B.17)}\\
&= \frac{1}{\delta}\left(-2R\sin\frac{\left(2i+1\right)\Delta\phi}{2}\sin\frac{\Delta\phi}{2}, 2R\cos\frac{\left(2i+1\right)\Delta\phi}{2}\sin\frac{\Delta\phi}{2}, \Delta z\right), && \text{(B.18)}\\
\mathbf{b}_i &= \frac{\mathbf{t}_{i-1}\times\mathbf{t}_i}{\left|\mathbf{t}_{i-1}\times\mathbf{t}_i\right|}, \quad \mathbf{n}_i = \mathbf{b}_i\times\mathbf{t}_i. && \text{(B.19)}
\end{aligned}$$

We can then calculate the bond angle

$$\begin{aligned}
\psi_i &= \arccos\left(\mathbf{t}_{i-1}\cdot\mathbf{t}_i\right) && \text{(B.20)}\\
&= \arccos\frac{4R^2\sin^2\frac{\Delta\phi}{2}\cos\Delta\phi + \Delta z^2}{\delta^2} && \text{(B.21)}\\
&= 1.56 = 89.5^\circ \approx \frac{\pi}{2}, && \text{(B.22)}
\end{aligned}$$

and the torsion angle

$$\begin{aligned}
\theta_i &= \arccos\left(\mathbf{b}_{i-1}\cdot\mathbf{b}_i\right) && \text{(B.23)}\\
&= \arccos\frac{\Delta z^2\cos\Delta\phi + R^2\sin^2\Delta\phi}{\Delta z^2 + R^2\sin^2\Delta\phi} && \text{(B.24)}\\
&= 0.879 = 50.4^\circ. && \text{(B.25)}
\end{aligned}$$

Since of the special value of bond angle, we can simplify the Frenet equations as following

$$\mathbf{b}_i = \frac{\mathbf{t}_{i-1} \times \mathbf{t}_i}{|\mathbf{t}_{i-1} \times \mathbf{t}_i|} = \frac{\mathbf{t}_{i-1} \times \mathbf{t}_i}{\sin \psi_i} \approx \mathbf{t}_{i-1} \times \mathbf{t}_i, \tag{B.26}$$

$$\mathbf{n}_i = \mathbf{b}_i \times \mathbf{t}_i \approx -\mathbf{t}_{i-1}, \tag{B.27}$$

$$\begin{pmatrix} \mathbf{t} \\ \mathbf{n} \\ \mathbf{b} \end{pmatrix}_i = \begin{pmatrix} 0 & \cos\theta & \sin\theta \\ -1 & 0 & 0 \\ 0 & -\sin\theta & \cos\theta \end{pmatrix}_i \begin{pmatrix} \mathbf{t} \\ \mathbf{n} \\ \mathbf{b} \end{pmatrix}_{i-1} . \tag{B.28}$$

And the chirality, defined by the sign of triple product $(\mathbf{t}_{i-1} \times \mathbf{t}_i) \cdot \mathbf{t}_{i+1} \approx \mathbf{b}_i \cdot \mathbf{t}_{i+1} \approx \sin\theta_{i+1}$, is essentially determined by $\sin\theta_{i+1}$.

Annexe C

Poisson bracket of bond angle and torsion angle

In order to study the dynamics of curve movement, Hamilton's equation of motion is a natural choice. For this purpose, we work out the Poisson bracket between bond angle and torsion angle (between curvature and torsion), within the framework of discrete (continuous) smoke ring model [30]. The result we get here will serve as the basics for the future work on a general model.

C.1 Poisson structure in continuous smoke ring equation

In 1972 Hasimoto proved [24] that the smoke ring equation

$$\frac{d\mathbf{r}}{dt} = \kappa \mathbf{b}, \tag{C.1}$$

is gauge equivalent to the nonlinear Schrodinger equation (NSE)

$$i\partial_t q = -\partial_s^2 q - \frac{1}{2}|q|^2 q, \tag{C.2}$$

where $q \equiv \kappa e^{i\int^s \tau ds'}$ is then called Hasimoto variable. The corresponding Poisson structure is

$$H = \int ds \left(|\partial_s q|^2 - \frac{1}{4}|q|^4\right), \tag{C.3}$$

$$\{q(s), q^*(s')\} = i\delta(s-s'). \tag{C.4}$$

In terms of curvature and torsion, the Hamiltonian translates into

$$H = \int ds \left((\partial_s \kappa)^2 + \kappa^2\tau^2 - \frac{1}{4}\kappa^4\right), \tag{C.5}$$

while the Poisson bracket changes to be

$$\begin{aligned}
\{q(s), q^*(s')\} &= \left\{\kappa e^{i\int^s \tau ds''}, \kappa e^{-i\int^{s'} \tau ds''}\right\} && \text{(C.6)}\\
&= \kappa\left\{e^{i\int^s \tau ds''}, \kappa e^{-i\int^{s'} \tau ds'}\right\} + e^{i\int^s \tau ds''}\left\{\kappa, \kappa e^{-i\int^{s'} \tau ds'}\right\} && \text{(C.7)}\\
&= \kappa e^{-i\int^{s'} \tau ds'}\left\{e^{i\int^s \tau ds''}, \kappa\right\} + \kappa e^{i\int^{s'} \tau ds'}\left\{\kappa, e^{-i\int^{s'} \tau ds''}\right\} && \text{(C.8)}\\
&= \kappa\left\{i\int^{s'} \tau ds'', \kappa\right\} + \kappa\left\{\kappa, -i\int^{s'} \tau ds''\right\} && \text{(C.9)}\\
&= 2i\kappa\left\{\kappa, -\int^{s'} \tau ds''\right\}. && \text{(C.10)}
\end{aligned}$$

So finally we get

$$\{\kappa(s), \tau(s')\} = \frac{1}{2\kappa(s)}\frac{\partial}{\partial s}\delta(s-s'). \tag{C.11}$$

There are two equivalent ways to calculate the equation of motion. One is to directly change the variable from Eq. (C.2). The other is to make use of Hamilton's dynamics based on both Eq. (C.5) and Eq. (C.11). Both give the same result

$$\begin{aligned}
\partial_t \kappa &= -2(\partial_s \kappa)\tau - \kappa\partial_s \tau, && \text{(C.12)}\\
\partial_t \tau &= \frac{\partial}{\partial s}\left(\frac{\partial_s^2\kappa - \kappa\tau^2 + \frac{1}{2}\kappa^3}{\kappa}\right). && \text{(C.13)}
\end{aligned}$$

Such kind of equivalence serves as the double check in the calculation. Similar strategy is used in the discrete case, as following.

C.2 Poisson structure in lattice Heisenberg model

Lattice Heisenberg model (LHM) is a natural discretization of smoke ring equation. Below we show its Poisson structure, in terms of both tangent vector and bond/torsion angles.

Combing both the discrete Frenet frame (DFF)

$$\begin{cases}
\mathbf{t}_i = \dfrac{\mathbf{r}_{i+1}-\mathbf{r}_i}{|\mathbf{r}_{i+1}-\mathbf{r}_i|} \equiv \dfrac{\mathbf{r}_{i+1}-\mathbf{r}_i}{\delta}, \mathbf{b}_i = \dfrac{\mathbf{t}_{i-1}\times\mathbf{t}_i}{|\mathbf{t}_{i-1}\times\mathbf{t}_i|}, \mathbf{n}_i = \mathbf{b}_i\times\mathbf{t}_i,\\
\begin{pmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{pmatrix}_{i+1} = \begin{pmatrix}\cos\psi & \cos\theta\sin\psi & \sin\theta\sin\psi\\ -\sin\psi & \cos\theta\cos\psi & \sin\theta\cos\psi\\ 0 & -\sin\theta & \cos\theta\end{pmatrix}_{i+1}\begin{pmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{pmatrix}_i,
\end{cases} \tag{C.14}$$

and the integrable LHM model [20]

$$\begin{aligned}
H &= -\frac{2}{\delta^2}\sum_i \log(1+\mathbf{t}_i\cdot\mathbf{t}_{i+1}), && \text{(C.15)}\\
\left\{\mathbf{t}_i^a, \mathbf{t}_j^b\right\} &= -\varepsilon^{abc}\mathbf{t}_i^c\delta_{ij}, && \text{(C.16)}
\end{aligned}$$

one can get the Heisenberg flow as following

$$\begin{aligned}
\frac{d\mathbf{t}_i}{dt} &= \{H, \mathbf{t}_i\} = -\mathbf{t}_i \times \frac{\partial H}{\partial \mathbf{t}_i} && \text{(C.17)}\\
&= \frac{2}{\delta^2}\left(\frac{\mathbf{t}_i \times \mathbf{t}_{i+1}}{1+\mathbf{t}_i\cdot\mathbf{t}_{i+1}} - \frac{\mathbf{t}_{i-1}\times\mathbf{t}_i}{1+\mathbf{t}_{i-1}\cdot\mathbf{t}_i}\right) && \text{(C.18)}\\
&= \frac{2}{\delta^2}\left(\frac{\sin\psi_{i+1}\mathbf{b}_{i+1}}{1+\cos\psi_{i+1}} - \frac{\sin\psi_i\mathbf{b}_i}{1+\cos\psi_i}\right) && \text{(C.19)}\\
&= \frac{2}{\delta^2}\left(\tan\frac{\psi_{i+1}}{2}\mathbf{b}_{i+1} - \tan\frac{\psi_i}{2}\mathbf{b}_i\right). && \text{(C.20)}
\end{aligned}$$

where the last step has used the relations

$$|\mathbf{t}_{i-1}\times\mathbf{t}_i|^2 = |\mathbf{t}_{i-1}|^2|\mathbf{t}_i|^2 - (\mathbf{t}_{i-1}\cdot\mathbf{t}_i)^2 = 1-\cos^2\psi_i = \sin^2\psi_i. \quad \text{(C.21)}$$

One feature of the model is its preservation of the closure condition of the curve

$$\sum_{i=1}^{N}\mathbf{t}_i = \mathbf{0}, \qquad \mathbf{t}_1 = \mathbf{t}_{N+1}, \quad \text{(C.22)}$$

$$\frac{d}{dt}\sum_{i=1}^{N}\mathbf{t}_i = \frac{2}{\delta^2}\sum_{i=1}^{N}\left(\tan\frac{\psi_{i+1}}{2}\mathbf{b}_{i+1} - \tan\frac{\psi_i}{2}\mathbf{b}_i\right) = 0. \quad \text{(C.23)}$$

The curve is then reconstructed as $\mathbf{r}_i = \mathbf{r}_{i-1} + \delta\mathbf{t}_{i-1}$. In the explicit way the flows on $\mathbf{r}_i$ reads

$$\frac{d\mathbf{r}_i}{dt} = \frac{2}{\delta}\frac{\mathbf{t}_{i-1}\times\mathbf{t}_i}{1+\mathbf{t}_{i-1}\cdot\mathbf{t}_i} = \frac{2}{\delta}\tan\frac{\psi_i}{2}\mathbf{b}_i, \quad \text{(C.24)}$$

which can be regarded as the discrete analogue of smoke ring equation, compared with Eq. (C.1).

C.2.1 Equation of motion for angles : by means of changing variable

We would like to change variable, in order to get the time dynamics of angles. To simplify the calculation, we rescale the time so that Eq. (C.18) is sent as $\frac{2}{\delta^2} \to 1$. For bond angles we get

$$\begin{aligned}
\frac{d(\cos\psi_i)}{dt} &= -\sin\psi_i\frac{d\psi_i}{dt}\\
&= \frac{d(\mathbf{t}_{i-1}\cdot\mathbf{t}_i)}{dt} = \frac{d\mathbf{t}_{i-1}}{dt}\cdot\mathbf{t}_i + \mathbf{t}_{i-1}\cdot\frac{d\mathbf{t}_i}{dt}\\
&= \left(\tan\frac{\psi_i}{2}\mathbf{b}_i - \tan\frac{\psi_{i-1}}{2}\mathbf{b}_{i-1}\right)\cdot\mathbf{t}_i + \mathbf{t}_{i-1}\cdot\left(\tan\frac{\psi_{i+1}}{2}\mathbf{b}_{i+1} - \tan\frac{\psi_i}{2}\mathbf{b}_i\right)\\
&= 0 - \tan\frac{\psi_{i-1}}{2}\sin\theta_i\sin\psi_i + \tan\frac{\psi_{i+1}}{2}\sin\theta_{i+1}\sin\psi_i - 0. && \text{(C.25)}
\end{aligned}$$

The last line has used the DFF

$$\begin{aligned}
\mathbf{t}_{i-1}\cdot\mathrm{b}_{i+1} &= \mathbf{t}_{i-1}\cdot(-\sin\theta_{i+1}\mathbf{n}_i + \cos\theta_{i+1}\mathbf{b}_i)\\
&= -\sin\theta_{i+1}\mathbf{t}_{i-1}\cdot\mathbf{n}_i\\
&= \sin\theta_{i+1}\sin\psi_i. && \text{(C.26)}
\end{aligned}$$

So we finally arrive at the equation for bond angles

$$\frac{d\psi_i}{dt} = \tan\frac{\psi_{i-1}}{2}\sin\theta_i - \tan\frac{\psi_{i+1}}{2}\sin\theta_{i+1}. \tag{C.27}$$

Taking similar strategy for torsion angles, we have

$$\frac{d\left(\cos\theta_i\right)}{dt} = -\sin\theta_i\frac{d\theta_i}{dt} = \frac{d\left(\mathbf{b}_{i-1}\cdot\mathbf{b}_i\right)}{dt} = \frac{d\mathbf{b}_{i-1}}{dt}\cdot\mathbf{b}_i + \mathbf{b}_{i-1}\cdot\frac{d\mathbf{b}_i}{dt}. \tag{C.28}$$

Here the calculation would be a little complicated. Firstly we compute the derivative of binormal vector

$$\begin{aligned}
\frac{d\mathbf{b}_i}{dt} &= \frac{d}{dt}\left(\frac{\mathbf{t}_{i-1}\times\mathbf{t}_i}{\sin\psi_i}\right)\\
&= -\frac{1}{\sin^2\psi_i}\cos\psi_i\frac{d\psi_i}{dt}\left(\mathbf{t}_{i-1}\times\mathbf{t}_i\right)\\
&\quad +\frac{1}{\sin\psi_i}\left(\frac{d\mathbf{t}_{i-1}}{dt}\times\mathbf{t}_i + \mathbf{t}_{i-1}\times\frac{d\mathbf{t}_i}{dt}\right)\\
&= -\cot\psi_i\frac{d\psi_i}{dt}\mathbf{b}_i + \csc\psi_i\left(\tan\frac{\psi_i}{2}\mathbf{n}_i - \tan\frac{\psi_{i-1}}{2}\mathbf{b}_{i-1}\times\mathbf{t}_i\right.\\
&\qquad\left. +\tan\frac{\psi_{i+1}}{2}\mathbf{t}_{i-1}\times\mathbf{b}_{i+1} - \tan\frac{\psi_i}{2}\mathbf{t}_{i-1}\times\mathbf{b}_i\right).
\end{aligned} \tag{C.29}$$

Then we have

$$\begin{aligned}
\frac{d\mathbf{b}_i}{dt}\cdot\mathbf{b}_{i-1} &= -\cot\psi_i\frac{d\psi_i}{dt}\cos\theta_i + \csc\psi_i\left(\tan\frac{\psi_i}{2}\cos\psi_i\sin\theta_i - 0+\right.\\
&\qquad\left.\tan\frac{\psi_{i+1}}{2}\left(\mathbf{t}_{i-1}\times\mathbf{b}_{i+1}\right)\cdot\mathbf{b}_{i-1} - \tan\frac{\psi_i}{2}\left(\mathbf{t}_{i-1}\times\mathbf{b}_i\right)\cdot\mathbf{b}_{i-1}\right).
\end{aligned} \tag{C.30}$$

The mixed product can be simplified as

$$\begin{aligned}
\left(\mathbf{t}_{i-1}\times\mathbf{b}_{i+1}\right)\cdot\mathbf{b}_{i-1} &= \left(\mathbf{b}_{i-1}\times\mathbf{t}_{i-1}\right)\cdot\mathbf{b}_{i+1} = \mathbf{n}_{i-1}\cdot\mathbf{b}_{i+1}\\
&= \mathbf{n}_{i-1}\cdot\left(-\sin\theta_{i+1}\mathbf{n}_i + \cos\theta_{i+1}\mathbf{b}_i\right)\\
&= -\sin\theta_{i+1}\cos\psi_i\cos\theta_i - \cos\theta_{i+1}\sin\theta_i, \qquad \text{(C.31)}\\
\left(\mathbf{t}_{i-1}\times\mathbf{b}_i\right)\cdot\mathbf{b}_{i-1} &= \left(\mathbf{b}_{i-1}\times\mathbf{t}_{i-1}\right)\cdot\mathbf{b}_i = \mathbf{n}_{i-1}\cdot\mathbf{b}_i = -\sin\theta_i. \qquad \text{(C.32)}
\end{aligned}$$

Combing the above three equations, we arrive at

$$\begin{aligned}
\frac{d\mathbf{b}_i}{dt}\cdot\mathbf{b}_{i-1} &= -\cot\psi_i\frac{d\psi_i}{dt}\cos\theta_i + \csc\psi_i\left(\tan\frac{\psi_i}{2}\left(\cos\psi_i+1\right)\sin\theta_i -\right.\\
&\qquad\left.\tan\frac{\psi_{i+1}}{2}\left(\cos\psi_i\sin\theta_{i+1}\cos\theta_i + \cos\theta_{i+1}\sin\theta_i\right)\right)\\
&= -\cot\psi_i\frac{d\psi_i}{dt}\cos\theta_i + \sin\theta_i\\
&\quad -\tan\frac{\psi_{i+1}}{2}\left(\cot\psi_i\sin\theta_{i+1}\cos\theta_i + \csc\psi_i\cos\theta_{i+1}\sin\theta_i\right).
\end{aligned} \tag{C.33}$$

The next step is to compute the other dot product

$$\begin{aligned}\mathbf{b}_i \cdot \frac{d\mathbf{b}_{i-1}}{dt} &= -\cot\psi_{i-1}\frac{d\psi_{i-1}}{dt}\cos\theta_i + \\ &\quad \csc\psi_{i-1}\left(-\tan\frac{\psi_{i-1}}{2}\sin\theta_i - \tan\frac{\psi_{i-2}}{2}\mathbf{b}_i\cdot(\mathbf{b}_{i-2}\times\mathbf{t}_{i-1})\right. \\ &\quad \left. +0 - \tan\frac{\psi_{i-1}}{2}\mathbf{b}_i\cdot(\mathbf{t}_{i-2}\times\mathbf{b}_{i-1})\right). \qquad \text{(C.34)}\end{aligned}$$

Again we need to compute the mixed product

$$\begin{aligned}\mathbf{b}_i\cdot(\mathbf{b}_{i-2}\times\mathbf{t}_{i-1}) &= (-\sin\theta_i\mathbf{n}_{i-1}+\cos\theta_i\mathbf{b}_{i-1})\cdot(\mathbf{b}_{i-2}\times\mathbf{t}_{i-1}) \\ &= -\sin\theta_i(\mathbf{t}_{i-1}\times\mathbf{n}_{i-1})\cdot\mathbf{b}_{i-2}+\cos\theta_i(\mathbf{t}_{i-1}\times\mathbf{b}_{i-1})\cdot\mathbf{b}_{i-2} \\ &= -\sin\theta_i\mathbf{b}_{i-1}\cdot\mathbf{b}_{i-2}-\cos\theta_i\mathbf{n}_{i-1}\cdot\mathbf{b}_{i-2} \\ &= -\sin\theta_i\cos\theta_{i-1}-\cos\theta_i\sin\theta_{i-1}\cos\psi_{i-1}, \qquad \text{(C.35)} \\ \mathbf{b}_i\cdot(\mathbf{t}_{i-2}\times\mathbf{b}_{i-1}) &= (-\sin\theta_i\mathbf{n}_{i-1}+\cos\theta_i\mathbf{b}_{i-1})\cdot(\mathbf{t}_{i-2}\times\mathbf{b}_{i-1}) \\ &= -\sin\theta_i(\mathbf{b}_{i-1}\times\mathbf{n}_{i-1})\cdot\mathbf{t}_{i-2}+\cos\theta_i(\mathbf{b}_{i-1}\times\mathbf{b}_{i-1})\cdot\mathbf{t}_{i-2} \\ &= \sin\theta_i\mathbf{t}_{i-1}\cdot\mathbf{t}_{i-2} \\ &= \sin\theta_i\cos\psi_{i-1} \qquad \text{(C.36)}\end{aligned}$$

Combing the above three equations, we arrive at

$$\begin{aligned}\mathbf{b}_i\cdot\frac{d\mathbf{b}_{i-1}}{dt} &= -\cot\psi_{i-1}\frac{d\psi_{i-1}}{dt}\cos\theta_i+\csc\psi_{i-1}\left(-\tan\frac{\psi_{i-1}}{2}(\cos\psi_{i-1}+1)\sin\theta_i+\right. \\ &\quad \left.\tan\frac{\psi_{i-2}}{2}(\cos\psi_{i-1}\sin\theta_{i-1}\cos\theta_i+\cos\theta_{i-1}\sin\theta_i)\right) \\ &= -\cot\psi_{i-1}\frac{d\psi_{i-1}}{dt}\cos\theta_i+\sin\theta_i \\ &\quad +\tan\frac{\psi_{i-2}}{2}(\cot\psi_{i-1}\sin\theta_{i-1}\cos\theta_i+\csc\psi_{i-1}\cos\theta_{i-1}\sin\theta_i). \qquad \text{(C.37)}\end{aligned}$$

This equation, together with Eq. (C.28) and Eq. (C.33), gives

$$\begin{aligned}\frac{d\theta_i}{dt} &= \cot\theta_i\left(\cot\psi_i\frac{d\psi_i}{dt}+\cot\psi_{i-1}\frac{d\psi_{i-1}}{dt}\right) \\ &\quad +\tan\frac{\psi_{i+1}}{2}(\cot\psi_i\sin\theta_{i+1}\cot\theta_i+\csc\psi_i\cos\theta_{i+1}) \\ &\quad -\tan\frac{\psi_{i-2}}{2}(\cot\psi_{i-1}\sin\theta_{i-1}\cot\theta_i+\csc\psi_{i-1}\cos\theta_{i-1}). \qquad \text{(C.38)}\end{aligned}$$

By means of Eq. (C.27), the first term can be rewritten as

$$\begin{aligned}\cot\theta_i\left(\cot\psi_i\frac{d\psi_i}{dt}+\cot\psi_{i-1}\frac{d\psi_{i-1}}{dt}\right) &= \cot\theta_i\left(\cot\psi_i\left(\tan\frac{\psi_{i-1}}{2}\sin\theta_i-\tan\frac{\psi_{i+1}}{2}\sin\theta_{i+1}\right)\right. \\ &+ \left.\cot\psi_{i-1}\left(\tan\frac{\psi_{i-2}}{2}\sin\theta_{i-1}-\tan\frac{\psi_i}{2}\sin\theta_i\right)\right) \qquad \text{(C.39)}\end{aligned}$$

This cancels two terms in Eq. (C.38) and finally we get the equation for torsion angles

$$\begin{aligned}\frac{d\theta_i}{dt} &= \cos\theta_i\left(\cot\psi_i\tan\frac{\psi_{i-1}}{2}-\cot\psi_{i-1}\tan\frac{\psi_i}{2}\right)\\ &\quad+\tan\frac{\psi_{i+1}}{2}\csc\psi_i\cos\theta_{i+1}-\tan\frac{\psi_{i-2}}{2}\csc\psi_{i-1}\cos\theta_{i-1}. \qquad \text{(C.40)}\end{aligned}$$

C.2.2 Equation of motion for angles : by means of Poisson bracket

Here we would like to start from the Poisson bracket in Eq. (C.16) to the Poisson brackets between bond/torsion angles. Since one bond angle involve two consecutive tangent vectors while one torsion angle involve three, there are nine nonvanishing brackets, i.e. $\{\psi_i,\psi_{i+1}\}$, $\{\psi_{i-2},\theta_i\}$, $\{\psi_{i-1},\theta_i\}$, $\{\psi_i,\theta_i\}$, $\{\psi_{i+1},\theta_i\}$, $\{\theta_{i-2},\theta_i\}$, $\{\theta_{i-1},\theta_i\}$, $\{\theta_{i+1},\theta_i\}$,$\{\theta_{i+2},\theta_i\}$. For the revalence with LHM, only the first five brackets are shown below. Since the calculation takes the same skill for each bracket, here only the details of computing $\{\psi_i,\psi_{i+1}\}$ is given, as following

$$\begin{aligned}\{\cos\psi_i,\cos\psi_{i+1}\} &= \sin\psi_i\sin\psi_{i+1}\{\psi_i,\psi_{i+1}\} &\text{(C.41)}\\ &= \{\mathbf{t}_i\cdot\mathbf{t}_{i-1},\mathbf{t}_i\cdot\mathbf{t}_{i+1}\} &\text{(C.42)}\\ &= \mathbf{t}_i\cdot\{\mathbf{t}_{i-1},\mathbf{t}_i\cdot\mathbf{t}_{i+1}\}+\{\mathbf{t}_i,\mathbf{t}_i\cdot\mathbf{t}_{i+1}\}\cdot\mathbf{t}_{i-1} &\text{(C.43)}\\ &= \mathbf{t}_i\cdot\left(\mathbf{t}_{i-1}\times\frac{\partial\left(\mathbf{t}_i\cdot\mathbf{t}_{i+1}\right)}{\partial\mathbf{t}_{i-1}}\right)+\left(\mathbf{t}_i\times\frac{\partial\left(\mathbf{t}_i\cdot\mathbf{t}_{i+1}\right)}{\partial\mathbf{t}_i}\right)\cdot\mathbf{t}_{i-1} &\text{(C.44)}\\ &= 0+\mathbf{t}_i\times\mathbf{t}_{i+1}\cdot\mathbf{t}_{i-1} &\text{(C.45)}\\ &= \mathbf{t}_{i-1}\times\mathbf{t}_i\cdot\mathbf{t}_{i+1} &\text{(C.46)}\\ &= \sin\psi_i\mathbf{b}_i\cdot\mathbf{t}_{i+1} &\text{(C.47)}\\ &= \sin\psi_i\sin\theta_{i+1}\sin\psi_{i+1}, &\text{(C.48)}\end{aligned}$$

So we get

$$\{\psi_i,\psi_{i+1}\}=\sin\theta_{i+1}. \qquad \text{(C.49)}$$

In the similar way, we can calculate the other brackets. The results are summarized as following

$$\begin{aligned}\{\psi_{i-2},\theta_i\} &= -\cos\theta_{i-1}\csc\psi_{i-1}, &\text{(C.50)}\\ \{\psi_{i-1},\theta_i\} &= \cot\frac{\psi_{i-1}}{2}+\cos\theta_i\cot\psi_i, &\text{(C.51)}\\ \{\psi_i,\theta_i\} &= -\cot\frac{\psi_i}{2}-\cos\theta_i\cot\psi_{i-1}, &\text{(C.52)}\\ \{\psi_{i+1},\theta_i\} &= \cos\theta_{i+1}\csc\psi_i, &\text{(C.53)}\\ \{\psi_i,\psi_{i+1}\} &= \sin\theta_{i+1}. &\text{(C.54)}\end{aligned}$$

At the same time, the Hamiltonian reads

$$\begin{aligned}H &= -\sum_i\log\left(1+\mathbf{t}_i\cdot\mathbf{t}_{i+1}\right) &\text{(C.55)}\\ &= -\sum_i\log\left(1+\cos\psi_i\right) &\text{(C.56)}\\ &= -2\sum_i\log\cos\frac{\psi_i}{2}. &\text{(C.57)}\end{aligned}$$

So we get

$$\frac{d\psi_i}{dt} = \{H, \psi_i\} \tag{C.58}$$
$$= -2\left\{\log\cos\frac{\psi_{i-1}}{2} + \log\cos\frac{\psi_{i+1}}{2}, \psi_i\right\} \tag{C.59}$$
$$= \tan\frac{\psi_{i-1}}{2}\sin\theta_i - \tan\frac{\psi_{i+1}}{2}\sin\theta_{i+1}. \tag{C.60}$$

$$\frac{d\theta_i}{dt} = \{H, \theta_i\} \tag{C.61}$$
$$= -2\left\{\log\cos\frac{\psi_{i-2}}{2} + \log\cos\frac{\psi_{i-1}}{2} + \log\cos\frac{\psi_i}{2} + \log\cos\frac{\psi_{i+1}}{2}, \theta_i\right\} \tag{C.62}$$
$$= \tan\frac{\psi_{i-2}}{2}\left(-\cos\theta_{i-1}\csc\psi_{i-1}\right) + \tan\frac{\psi_{i-1}}{2}\left(\cot\frac{\psi_{i-1}}{2} + \cos\theta_i\cot\psi_i\right) \tag{C.63}$$
$$+ \tan\frac{\psi_i}{2}\left(-\cot\frac{\psi_i}{2} + \cos\theta_i\cot\psi_{i-1}\right) + \tan\frac{\psi_{i+1}}{2}\cos\theta_{i+1}\csc\psi_i \tag{C.64}$$
$$= \cos\theta_i\left(\cot\psi_i\tan\frac{\psi_{i-1}}{2} - \cot\psi_{i-1}\tan\frac{\psi_i}{2}\right) \tag{C.65}$$
$$+ \tan\frac{\psi_{i+1}}{2}\cos\theta_{i+1}\csc\psi_i - \tan\frac{\psi_{i-2}}{2}\csc\psi_{i-1}\cos\theta_{i-1}. \tag{C.66}$$

Both equations agree with Eq. (C.27) and Eq. (C.40). Finally it is straightforward to check the Jacobi identity :

$$\{\psi_{i-1}, \{\theta_i, \psi_i\}\} + \{\theta_i, \{\psi_i, \psi_{i-1}\}\} + \{\psi_i, \{\psi_{i-1}, \theta_i\}\} \tag{C.67}$$
$$= \left\{\psi_{i-1}, \cot\frac{\psi_i}{2} + \cos\theta_i\cot\psi_{i-1}\right\} + \{\theta_i, -\sin\theta_i\} \tag{C.68}$$
$$+ \left\{\psi_i, \cot\frac{\psi_{i-1}}{2} + \cos\theta_i\cot\psi_i\right\} \tag{C.69}$$
$$= -\csc^2\frac{\psi_i}{2}\frac{1}{2}\sin\theta_i + \cot\psi_{i-1}\left(-\sin\theta_i\right)\left(\cot\frac{\psi_{i-1}}{2} + \cos\theta_i\cot\psi_i\right) + 0 \tag{C.70}$$
$$-\csc^2\frac{\psi_{i-1}}{2}\frac{1}{2}\left(-\sin\theta_i\right) + \cot\psi_i\left(-\sin\theta_i\right)\left(-\cot\frac{\psi_i}{2} - \cos\theta_i\cot\psi_{i-1}\right) \tag{C.71}$$
$$= \sin\theta_i\left(-\frac{1}{2}\csc^2\frac{\psi_i}{2} - \cot\psi_{i-1}\cot\frac{\psi_{i-1}}{2} + \frac{1}{2}\csc^2\frac{\psi_{i-1}}{2} + \cot\psi_i\cot\frac{\psi_i}{2}\right) \tag{C.72}$$
$$= \frac{1}{2}\sin\theta_i\left(\csc^2\frac{\psi_i}{2}\left(\cos\psi_i - 1\right) - \csc^2\frac{\psi_{i-1}}{2}\left(\cos\psi_{i-1} - 1\right)\right) \tag{C.73}$$
$$= 0. \tag{C.74}$$

C.2.3 Equivalence between lattice Heisenberg model and lattice nonlinear Schrodinger model

Define the discrete Hasimoto variable as

$$q_i = \tan\frac{\psi_i}{2}e^{i\sigma_i}, \sigma_i = \frac{1}{2}\left(\sum_{k=1}^{i}\theta_k - \sum_{k=i+1}^{N}\theta_k\right), \tag{C.75}$$

where the periodic boundary condition has been applied, i.e. $\theta_1 \equiv \theta_{N+1}$; N is the total number of the vertices. From Eq. (C.40), we notice that the sum

$$\begin{aligned}
\sum_k \frac{d\theta_k}{dt} &= \sum_k \cos\theta_k \left(\cot\psi_k \tan\frac{\psi_{k-1}}{2} - \cot\psi_{k-1}\tan\frac{\psi_k}{2} \right. \\
&\quad \left. + \csc\psi_{k-1}\tan\frac{\psi_k}{2} - \csc\psi_k \tan\frac{\psi_{k-1}}{2} \right) && \text{(C.76)} \\
&= \sum_k \cos\theta_k \left((\cot\psi_k - \csc\psi_k)\tan\frac{\psi_{k-1}}{2} - \right. \\
&\quad \left. (\cot\psi_{k-1} - \csc\psi_{k-1})\tan\frac{\psi_k}{2} \right) && \text{(C.77)} \\
&= \sum_k \cos\theta_k \left(\tan\frac{\psi_k}{2}\tan\frac{\psi_{k-1}}{2} - \tan\frac{\psi_{k-1}}{2}\tan\frac{\psi_k}{2} \right) && \text{(C.78)} \\
&= 0. && \text{(C.79)}
\end{aligned}$$

As a result, the derivative of $\frac{d\sigma_i}{dt}$ has only boundary contribution, i.e.

$$\begin{aligned}
\frac{d\sigma_i}{dt} &= \frac{1}{2}\left(\sum_{k=1}^{i} \frac{d\theta_k}{dt} - \sum_{k=i+1}^{N} \frac{d\theta_k}{dt} \right) && \text{(C.80)} \\
&= \cos\theta_{i+1}\csc\psi_i \tan\frac{\psi_{i+1}}{2} + \cos\theta_i \csc\psi_i \tan\frac{\psi_{i-1}}{2}. && \text{(C.81)}
\end{aligned}$$

Here the contributions from both summation cancel the factor $1/2$. The above equation further implies

$$2\tan\frac{\psi_i}{2}\frac{d\sigma_i}{dt} = \sec^2\frac{\psi_i}{2}\left(\tan\frac{\psi_{i+1}}{2}\cos\theta_{i+1} + \tan\frac{\psi_{i-1}}{2}\cos\theta_i \right). \tag{C.82}$$

On the other hand we have

$$\frac{dq_i}{dt} = \frac{1}{2}\sec^2\frac{\psi_i}{2} e^{i\sigma_i}\frac{d\psi_i}{dt} + \tan\frac{\psi_i}{2} e^{i\sigma_i} i\frac{d\sigma_i}{dt}. \tag{C.83}$$

$$\left(1 + |q_i|^2\right)(q_{i+1} + q_{i-1}) = \sec^2\frac{\psi_i}{2}\left(\tan\frac{\psi_{i+1}}{2} e^{i\theta_{i+1}} + \tan\frac{\psi_{i-1}}{2} e^{-i\theta_i} \right) e^{i\sigma_i}. \tag{C.84}$$

Combining the above three equations and Eq. (C.27) we get

$$2i\frac{dq_i}{dt} = -\left(1 + |q_i|^2\right)(q_{i+1} + q_{i-1}). \tag{C.85}$$

The factor 2 can be absorbed into the time. By replacing q_i with $q_i e^{it}$, we arrive at the lattice nonlinear Schrodinger equation

$$i\frac{dq_i}{dt} = -(q_{i+1} - 2q_i + q_{i-1}) - |q_i|^2 (q_{i+1} + q_{i-1}). \tag{C.86}$$

Bibliographie

[1] Shmygelska A, *Search for folding nuclei in native protein structures*, Bioinformatics **21** (2005), 394–402.

[2] Bissan Al-Lazikani, Arthur M. Lesk, and Cyrus Chothia, *Standard conformations for the canonical structures of immunoglobulins*, J. Mol. Biol. **273** (1997), 927–948.

[3] C. B. Anfinsen, *The limited digestion of ribonuclease with pepsin*, J. Biol. Chem **221** (1956), 405–412.

[4] R. L. Baldwin and G. D. Rose, *Is protein folding hierarchic ? 1. local structure and peptide folding*, Trends Biochem. Sci. **24** (1999), 26–33.

[5] D. Bashford, F. E. Cohen, M. Karplus, I. D. Kuntz, and D. L. Weaver, *Diffusion-collision model for the folding kinetics of myoglobin*, Proteins **4** (1988), 211–227.

[6] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *The protein data bank*, Nucl. Acids Res. **28** (2000), 235–242.

[7] J. Bryngelson, J. N. Onuchic, J. N. Socci, and P. G. Wolynes, *Funnels, pathways, and the energy landscape of protein folding : a synthesis*, Proteins **21** (1995), 167–195.

[8] R. E. Burton, J. K. Myers, and T. G. Oas, *Protein folding dynamics : quantitative comparison between theory and experiment*, Biochemistry **37** (1998), 5337–43.

[9] H. S. Chan and K. A. Dill, *Protein folding in the landscape perspective : Chevron plots and non-arrhenius kinetics*, Proteins **30** (1998), 2–33.

[10] C. Chothia, *Proteins. one thousand families for the molecular biologist*, Nature **357** (1992), 543–544.

[11] Ulf H. Danielsson, Martin Lundgren, and Antti J. Niemi, *A gauge field theory of chirally folded homopolymers with applications to folded proteins*, Phys. Rev. E **82** (2010), 021910.

[12] T. Dauxois and M. Peyrard, *Physics of solitons*, Cambridge University Press, England, 2006.

[13] A. S. Davydov, *The theory of contraction of proteins under their excitation*, J. Theor. Bio. **38** (1973), 559–569.

[14] ———, *Solitons and energy transfer along protein molecules*, J. Theor. Bio. **66** (1977), 379–387.

[15] K. A. Dill, *Dominant forces in protein folding*, Biochemistry **29** (1990), 7133–55.

[16] K. A. Dill, K. M. Fiebig, and H. S. Chan, *Cooperativity in protein-folding kinetics*, Proc. Natl. Acad. Sci. USA **90** (1993), 1942–1946.

[17] Ken A. Dill, *Polymer principles and protein folding*, Protein Sci. **8** (1999), 1166–1180.

[18] Ken A. Dill, S. Banu Ozkan, M. Scott Shell, and Thomas R. Weikl, *The protein folding problem*, Annu. Rev. Biophys. **37** (2008), 289–316.

[19] C. M. Dobson and M. Karplus, *The fundamentals of protein folding : bringing together theory and experiment*, Curr. Opin. Struct. Biol. **9** (1999), 92–101.

[20] L. D. Faddeev and L. A. Takhtajan, *Hamiltonian methods in the theory of solitons*, ch. I, pp. 296–305, 1987.

[21] A. Fersht, *Structure and mechanism in protein science : A guide to enzyme catalysis and protein folding*, Freeman, New York, 1999.

[22] K. M. Fiebig and K. A. Dill, *Protein core assembly processes*, J. Chem. Phys. **98** (1993), 3475–3487.

[23] P. J. Flory, *Statistical mechanics of chain molecules*, Wiley, New York, 1969.

[24] H. Hasimoto, *A soliton on a vortex filament*, J. Fluid. Mech. **51** (1972), 477–485.

[25] N Haspel, CJ Tsai, H Wolfson, and R Nussinov, *Reducing the computational complexity of protein folding via fragment folding and assembly*, Protein Sci. **12** (2003), 1177–87.

[26] A. Hausrath and A. Goriely, *Protein architectures predicted by a continuum representation of fold space*, Protein Sci. **15** (2006), 753–760.

[27] A. P. Heath, L. E. Kavraki, and Clementi C., *From coarse-grain to all-atom : toward multiscale analysis of protein landscapes*, Proteins **68** (2007), 646–661.

[28] M. Herrmann, *Heteroclinic standing waves in defocusing dnls equations : variational approach via energy minimization*, Applicable Analysis **89** (2010), 1591–1602.

[29] J Hockenmaier, AK Joshi, and KA Dill, *Routes are trees : the parsing perspective on protein folding*, Proteins **66** (2006), 1–15.

[30] S. Hu and Antti J. Niemi, *to appear*.

[31] F Huang, S Sato, TD Sharpe, L Ying, and AR Fersht, *Distinguishing between cooperative and unimodal downhill protein folding*, Proc. Natl. Acad. Sci. USA **104** (2007), 123–27.

[32] S. E. Jackson, *How do small single-domain proteins fold ?*, Fold. Des. **3** (1998), R81–91.

[33] Peng Jian and Jinbo Xu, *A multiple-template approach to protein threading*, Proteins **79** (2011), 1930Ð1939.

[34] D. T. Jones, W. R. Taylor, and J. M. Thornton, *A new approach to protein fold recognition*, Nature **358** (1992), 86–89.

[35] M. Karplus and D. L. Weaver, *Protein folding dynamics : the diffusion-collision model and experimental data*, Protein Sci. **3** (1994), 650–68.

[36] Yevgeny Kats, David A. Kessler, and Yitzhak Rabin, *Frenet algorithm for simulations of fluctuating continuous elastic filaments*, Phys. Rev. E **65** (2002), 020801(R).

[37] S. K. Kearsley, *On the orthogonal transformation used for structural comparisons*, Acta Crystallogr. A **45** (1989), 208–210.

[38] Andrey Krokhotin, Martin Lundgren, and Antti J. Niemi, *Soliton driven relaxation dynamics and universality in protein collapse*, arXiv :1111.2028 (2011).

[39] C. Levinthal, *Are there pathways for protein folding ?*, J. Chim. Phys. **65** (1968), 44.

[40] MO Lindberg and M Oliveberg, *Malleability of protein folding pathways : a simple reason for complex behaviour*, Curr. Opin. Struct. Biol. **17** (2007), 21–29.

[41] H Maity, M Maity, MMG Krishna, L Mayne, and SW Englander, *Protein folding : the stepwise assembly of foldon units*, Proc. Natl. Acad. Sci. **102** (2005), 4741–4746.

[42] N. Manton and P. Sutcliffe, *Topological solitons*, Cambridge University Press, England, 2004.

[43] A. E. Mirsky and L. Pauling, *On the structure of native, denatured, and coagulated proteins.*, Proc. Natl. Acad. Sci. USA **22** (1936), 439–447.

[44] A. L. Morris, M. W. MacArthur, E. G. Hutchinson, and J. M. Thornton, *Stereochemical quality of protein structure coordinates*, Proteins **12** (1992), 345–364.

[45] J. K. Myers and T. G. Oas, *Preorganized secondary structure as an important determinant of fast protein folding*, Nature Struct. Biol. **8** (2001), 552–8.

[46] Antti J. Niemi, *Phases of bosonic strings and two dimensional gauge theories*, Phys. Rev. D **67** (2003), 106004.

[47] N.C. Pace, *Measuring and increasing protein stability*, Trends in Biotech. **8** (1990), 93–98.

[48] Gregory A Petsko and Dagmar Ringe, *Protein structure and function*, New Science Press Ltd, London, 2004.

[49] O. B. Ptitsyn, *How the molten globule became*, Trends Biochem. Sci. **20** (1995), 376–9.

[50] S. Rackovsky and H.A. Scheraga, *Differential geometry and polymer conformations 1. Comparison of protein conformations*, Macromolecules **11** (1978), 1168–1174.

[51] ———, *Differential geometry and polymer conformations 2. Development of a conformational distance function*, Macromolecules **13** (1980), 1440–1453.

[52] ———, *Differential geometry and polymer conformations 3. Single-site and nearest-neighbor distribution and nucleation of protein folding*, Macromolecules **14** (1981), 1259–1269.

[53] Harold A. Scheraga, Mey Khalili, and Adam Liwo, *Protein-Folding Dynamics : Overview of Molecular Simulation Techniques*, Annu. Rev. Phys. Chem. **58** (2007), 57–83.

[54] J. Scott Russell, *Report on Waves*, Report of the fourteenth meeting of the British Association for the Advancement of Science.

[55] T. R. Sosnick, L. Mayne, and S. W. Englander, *Molecular collapse : The ratelimiting step in two-state cytochrome c folding*, Proteins **24** (1996), 413–426.

[56] Valentina Tozzini, *Coarse-grained models for proteins*, Curr. Opin. Struct. Biol. **15** (2005), 144–150.

[57] CJ Tsai, JV Jr Maizel, and R Nussinov, *Anatomy of protein structures : visualizing how a one-dimensional protein chain folds into a three-dimensional shape*, Proc. Natl. Acad. Sci. USA **97** (2000), 4741–4746.

[58] VA Voelz and KA Dill, *Exploring zipping and assembly as a protein folding principle*, Proteins **66** (2006), 877–88.

[59] TR Weikl, *Loop-closure events during protein folding : rationalizing the shape of ?-value distributions*, Proteins **60** (2005), 701–11.

[60] TR Weikl and KA Dill, *Folding kinetics of two-state proteins : effect of circularization, permutation, and crosslinks*, J. Mol. Biol. **332** (2003), 953–63.

[61] ———, *Folding rates and low-entropy-loss routes of two-state proteins*, J. Mol. Biol. **329** (2003), 585–98.

[62] TR Weikl, M Palassini, and KA Dill, *Cooperativity in two-state protein folding kinetics*, Protein Sci. **13** (2004), 822–29.

[63] R. Zwanzig, A. Szabo, and B. Bagchi, *Levinthal's paradox*, Proc. Natl. Acad. Sci. USA **89** (1992), 20–22.