Deterministic Global Optimization and Torsion Angle Dynamics for Molecular Structure Prediction

J. L. Klepeis and C.A. Floudas[1]

Department of Chemical Engineering, Princeton University, Princeton, N.J. 08544-5263

Abstract The problem of protein folding has become the subject of intense theoretical and experimental study over the past few decades. This work presents a new method for protein structure prediction using distance and dihedral angle restraints derived from NMR data. Both the formulation and solution approach differ substantially from traditional molecular structure prediction techniques. The traditional formulation is recast as a constrained global optimization problem whose solution is obtained through the use of the αBB algorithm, a deterministic global optimization approach suitable for nonconvex constrained problems. To enhance the efficiency of this method, torsion angle dynamics is introduced as an integral part of the solution approach. The proposed algorithm is tested on the Compstatin structure prediction problem.

Introduction

To effectively determine protein function it is important to predict the three dimensional structure of the macromolecule. Over the last several decades a number of experimental and theoretical approaches have been developed and refined in order to achieve this goal. One experimental technique, NMR (nuclear magnetic resonance) spectroscopy, is based on solution measurements of the system, and several key developments, including multidimensional NMR, have resulted in the ability to determine solution structures for proteins consisting of over 200 residues.

This work focuses on the development of a novel approach for protein structure prediction via experimental NMR restraints. Traditionally, the protein folding global optimization problem involves a progression of unconstrained minimizations. However, the introduction of experimentally derived or artificial restraints can be used to recast the fundamental protein folding problem as a constrained global optimization problem. The constraints, through reduction of the feasible search space, serve two important purposes : 1) attempt to correct any deficiencies of the energy model, and 2) focus the efforts of the global optimization algorithm.

The proposed constrained formulation differs from traditional NMR approaches in several fundamental ways [14]. First, the energy model is represented by a detailed full atom force field, rather than simplified nonbonded potential terms, which should make the approach especially effective when the number of NMR restraints per residue decreases. In addition, traditional solution approaches apply target function distance geometry or simulated annealing to unconstrained problem formulations in which restraints are represented by penalty function approximations. The solution of the constrained formulation requires the use of constrained local optimization solvers and an overall global optimization framework for nonlinearly constrained problems. This is accomplished through the application of the αBB deterministic global optimization approach [1, 2, 3].

Because the location of the global minimum relies on effectively solving constrained local optimization problems, convergence to the global minimum can be enhanced by consistently identifying low energy solutions. These observations illustrate the need for reliably locating low energy feasible points for initializing the constrained local optimization routine. Along these lines, torsion angle dynamics (TAD) has also been implemented within the context of the global optimization framework.

Theory

Basic data obtained from NMR studies consist of distance and torsion angle restraints. Once resonances have been assigned, nuclear Overhauser effect (NOE) contacts are selected and their intensities are used to calculate interproton distances. Information on torsion angles are based on the measurement of coupling constants and analysis of proton chemical shifts. Together, this information is used to formulate a nonlinear optimization problem, the solution of which should provide the correct protein structure. Typically, a hybrid energy function of the following form

[1]Author to whom all correspondence should be addressed; Tel.: (609) 258-4595; Fax: (609) 258-0211; email: floudas@titan.princeton.edu

is employed:

$$E = E_{\text{forcefield}} + W_{\text{nmr}} E_{\text{nmr}}. \tag{1}$$

The energy, E, specified by this target function includes a chemical description of the protein conformation through the use of a force field, $E_{\text{forcefield}}$. The force field potentials are generally much simpler representations of all atom force fields. The distance and dihedral angle restraints appear as weighted penalty, E_{nmr}, terms that should be driven to zero.

The second term of (1) can be rewritten as :

$$E_{\text{nmr}} = E_{\text{distance}} + E_{\text{dihedral}}. \tag{2}$$

Here E_{distance} and E_{dihedral} represent the violation energies based on the distance and dihedral angle restraints, respectively. These functions can take several forms, although a simple square well potential is typically used.

The force field energy term, $E_{\text{forcefield}}$ of Equation (1), models the nonbonded interactions of the protein, which can consist of simple or more detailed energy functions. In practice, when considering NMR restraints, force field terms are often simplified to include only geometric energy terms such as quartic Van der Waals repulsions. Such functions neglect rigorous modeling of energetic terms in order to ensure that experimental distance violations are minimized. A basic representation for this target function is :

$$E_{\text{S}} = E_{\text{distance}} + E_{\text{dihedral}}. \tag{3}$$

Here the E_{distance} function includes additional distance restraints to avoid van der Waals contacts.

In this work, a detailed modeling approach is proposed by using the ECEPP/3 force field [17]. For this force field, it is assumed that the covalent bond lengths and bond angles are fixed at their equilibrium values, so that the conformation is only a function of the independent torsional angles of the system. The total force field energy, $E_{\text{forcefield}}$, is calculated as the sum of electrostatic, nonbonded, hydrogen bonded, and torsional contributions. The main energy contributions (electrostatic, nonbonded, hydrogen bonded) are computed as the sum of terms for each atom pair whose interatomic distance is a function of at least one dihedral angle. This detailed modeling approach has as objective function :

$$E_{\text{D}} = E_{\text{distance}} + E_{\text{dihedral}} + E_{\text{ECEPP/3}}. \tag{4}$$

Global Optimization

The determination of a three dimensional protein structure defines an optimization problem in which the objective function may correspond to one of the target functions outlined above. A standard procedure for addressing this global optimization problem consists of a combination of simulated annealing and molecular or torsional angle dynamics [6]. Generally, multiple initial conformers are generated and optimized to provide a set of acceptable structures, which are then used to characterize the system. The simulated annealing protocol is incorporated in an attempt to reduce trapping in local minimum energy wells.

However, the minimization of complex target functions necessitates the use of rigorous global optimization techniques. Through the use of the constrained optimization approach, the dihedral angle bounds are implicitly included as box constraints. Furthermore, distance restraints are treated explicitly. This reformulation removes both E_{dihedral} and E_{distance} from the target function, leaving only $E_{\text{forcefield}}$:

$$\begin{aligned} \min_{\phi} \quad & E_{\text{ECEPP/3}} \\ \text{st} \quad & E_l^{\text{distance}}(\phi) \leq E_l^{\text{ref}} \quad l = 1, \ldots, N_{\text{CON}} \\ & \phi_i^L \leq \phi_i \leq \phi_i^U, \quad i = 1, \ldots, N_\phi. \end{aligned} \tag{5}$$

Here $i = 1, \ldots, N_\phi$ corresponds to the set of dihedral angles, ϕ_i, with ϕ_i^L and ϕ_i^U representing lower and upper bounds on these dihedral angles. E_l^{ref} is a reference parameter for the N_{CON} constraints. In general, the lower and upper bounds for these variables are set to -π and π, although appropriate bounds derived from NMR data are also suitable. The detailed atomistic-level energy function produces a multidimensional surface with an astronomically large number of local minima. To overcome these difficulties, the αBB global optimization approach [1, 2, 3] has been extended to identifying global minimum energy conformations of peptides. The algorithm has been shown to be successful for isolated peptide systems using the ECEPP/3 potential energy model [4, 15], and including several solvation models [12, 13]. αBB based global optimization techniques have also been applied to NMR type structure prediction problems [14, 20].

The αBB global optimization approach effectively brackets the global minimum by developing converging sequences of lower and upper bounds. These bounds are refined by iteratively partitioning the initial domain. Upper bounds on the global minimum are obtained by local minimizations of the original nonconvex problem. Lower bounds belong to the set of solutions of the convex lower bounding problems, which are constructed by augmenting the objective and constraint functions through the addition of separable quadratic terms.

Once solutions for the upper and lower bounding problems have been established, the next step is to modify these problems for the next iteration. This is accomplished by successively partitioning the initial domain into smaller subdomains. In order to ensure non–decreasing lower bounds, the hyper–rectangle to be bisected is chosen by selecting the region which contains the infimum of the minima of lower bounds. A non–increasing sequence for the upper bound is found by solving the nonconvex problem locally and selecting it to be the minimum over all the previously recorded upper bounds. If the single minimum of the convex relaxation for any hyper–rectangle is greater than the current upper bound, this hyper–rectangle can be discarded because the global minimum cannot lie within this subdomain (fathoming step). A comprehensive treatment of theory, methods and applications of deterministic global optimization can be found in [9].

Torsion Angle Dynamics

Standard unconstrained molecular dynamics simulations have been used extensively to model the folding and unfolding of protein systems [8, 7]. In addition, several methods for NMR structure calculation have been based on molecular dynamics in Cartesian space [6]. Torsion angle dynamics differs from traditional molecular dynamics in that bond lengths and bond angles are fixed at equilibrium values, thereby allowing for a transformation from the Cartesian to the internal coordinate system. The constraints on the systems also dampen high frequency motions, which permits the use of longer time steps during the numerical integration of the equations of motion. The use of TAD in place of conventional MD has been found to improve the efficiency of NMR structure prediction [10, 18].

A major disadvantage for employing TAD in place of Cartesian MD is that the equations of motion become much more complex for the constrained system. To avoid the potentially prohibitive computational cost required for the solution of the equations of motion, a fast recursive algorithm, which scales linearly with the number of torsion angles, was implemented. The algorithm is based on spatial operator algebra which has been used to simulate the dynamics of astronautical and robotic equipment [11].

The algorithm solves for the torsional accelerations, $\ddot{\boldsymbol{\theta}}$:

$$M(\boldsymbol{\theta})\ddot{\boldsymbol{\theta}} \;+\; C(\boldsymbol{\theta},\dot{\boldsymbol{\theta}}) \;=\; 0. \qquad (6)$$

In this equation M is an $N \times N$ nonlinear mass matrix and C is the N dimensional vector of velocity dependent (Coriolis and other) forces. $\boldsymbol{\theta}$, $\dot{\boldsymbol{\theta}}$ and $\ddot{\boldsymbol{\theta}}$ represent the torsional position, velocities and accelerations, respectively. The ability to calculate the accelerations recursively relies on the chainlike structure of the protein, in which each node of the chain represents a rigid body. These rigid bodies consist of one atom or a cluster of atoms whose relative positions are fixed.

The TAD is carried out using simulated annealing, with temperature control provided by coupling to an external bath [5]. This coupling provides a means for forcing or damping the torsional velocities using the following scaling factor at time t:

$$f_T = \sqrt{1 - \frac{1}{\beta} + \frac{T_o}{\beta T(t)}}. \qquad (7)$$

In this equation, β is a force constant, while T_o is the bath temperature and $T(t)$ is the actual temperature. Once torsional velocities have been determined, the accelerations, $\ddot{\boldsymbol{\theta}}$, can be calculated using the recursive algorithm. A basic leap-frog technique is then employed to calculate velocities at the half time-step, which can be used to calculate torsional positions, $\boldsymbol{\theta}$, and new estimated velocities at the full time step.

Computational Study

The global optimization algorithm was tested on Compstatin, a synthetic 13-residue (ICVVQDWGHHRCT) cyclic peptide (disulfide bridge between the Cys^2 and Cys^{12} residues) that binds to C3 (third component of complement) and inhibits complement activation [19]. Two-dimensional NMR techniques [16] yield a total of 30 backbone sequential (including H$^\beta$ – backbone), 23 medium and long range (including disulfide) and 82 intra-residue NOE restraints. In addition, 7 ϕ angle and 2 χ_1 angle dihedral restraints are available. In previous work [16], traditional distance geometry–simulated annealing protocol was utilized to minimize the associated target function in the Cartesian coordinate space using the program X-PLOR [6]. NOE distance and dihedral angle restraints were modeled using a quadratic square well potential, while van der Waals overlaps were prevented through the use of a simple quartic potential function.

The NMR refinement protocols resulted in a family of 21 structures with similar geometries for the Gln^5–Gly^8 region. A representative structure was obtained by averaging the coordinates of the individually refined structures and then subjecting this structure to further refinement to release geometric strain produced by the averaging process. The formation of a type I β-turn was identified as a common characteristic for these structures.

Local Minimization
The consistency of the ensemble of Compstatin solution structures was determined by evaluating distance restraints for each of the original 21 structures (accession number 1a1p at the Brookhaven Protein Data Bank, http://www.pdb.bnl.gov), as well as the average Compstatin conformation. In considering distance restraints, only backbone sequential and medium/long range NOEs were considered. That is, the 82 intra-residue restraints were neglected since they are less likely to effect the overall fold of the Compstatin peptide. This results in a total of 52 restraints, with an additional restraint on the distance between the sulfur atoms forming the disulfide bridge.

The results of the analysis indicate that the average structure ($\overline{Compstatin}$) possesses the third largest violation energy, whereas the smallest value is provided by structure 8 ($< Compstatin >_8$). These quantities provide a range of comparison for violation energies and were used to set the constraint parameter, E^{ref}, to 200 kcal/mol. This value is chosen so that the sum of the violation energies will necessarily result in an improvement over the violation energy for the average Compstatin structure.

Due to the relatively large distance violations and energies obtained after transformation of PDB to ECEPP/3 structures, the 22 structures were then subjected to local minimization. In all cases, the corresponding violation energy reached the upper bound value of 200 kcal/mol. The corresponding total distance violations increased, with an average value of 6.766 $\AA$. The smallest distance violation (5.873 $\AA$) was reported for structure number 10 ($< Compstatin >_{10}^{Local}$), whereas the corresponding energy for this structure (-41.685 kcal/mol) was only slightly above the average energy of -47.75 kcal/mol. The lowest energy structures (-71.613 for $< Compstatin >_2^{Local}$, -68.704 kcal/mol for $< Compstatin >_{21}^{Local}$, -67.653 kcal/mol for $< Compstatin >_9^{Local}$) provided above average values for total distance violation (6.963 $\AA$, 6.832 $\AA$, 7.120 $\AA$, respectively). In addition, the conformation obtained from the average Compstatin structure ($\overline{Compstatin}$) exhibited near average values for energy (-52.283 kcal/mol) and total distance violations (6.392 $\AA$).

The structural characteristics of these locally minimized structures were quantified using RMSD (root mean squared deviation) calculations. For the original PDB structures, comparison with the average Compstatin structure provided RMSD values between $1 - 2\AA$ for only backbone atoms. As expected, these structures possess common structural features. However, when comparing original PDB structures and their locally minimized counterparts, most RMSD values are larger than 2 $\AA$, indicating that significant conformational changes occur during local minimization. This is due to both the reduced set of NOE restraints in the constraint function and the role of the detailed energy force field. In contrast, the RMSD values for the β-turn region remain consistently low when comparing the original PDB structures to their locally minimized counterparts. These results indicate that the β-turn is a conserved structural feature.

Global Optimization without TAD
The constrained global optimization approach was first applied to Compstatin structure prediction without the use of TAD. A subset of 26 (all ϕ and ψ) torsion angles, from a total of 73, were treated globally, while the remaining were allowed to vary locally. As was the case for local minimization, the same set of restraints were used to formulate the nonlinear constraint. The lowest energy structure satisfying the constraint function provided an ECEPP/3 energy of -85.71 kcal/mol, an energy value more than 15 kcal/mol lower than those values provided by local minimization. The global minimization required approximately 40 CPU hours on a HP C160. The total distance violation equaled 6.690 $\AA$ which is near the average distance violation for the local minima.

RMSD calculations were performed to again quantify the structural differences between the global minimum energy structure and the other Compstatin structures. When comparing full backbone RMSD values, the $< Compstatin >_9^{Local}$, $< Compstatin >_{21}^{Local}$, $< Compstatin >_{19}^{Local}$ and $< Compstatin >_{17}^{Local}$ provide the best agreement with the global minimum energy structure. These structures also correspond to four of the lowest energy local minimum, indicating that some of the lowest energy conformers exhibit similar backbone structural characteristics. In contrast, the lowest energy local minimum, $< Compstatin >_2^{Local}$, is less similar to the global minimum energy structure. For the β-turn segment, the correlation between low RMSD values and low energy local minima does not exist. This observation, coupled with the relatively low RMSD values between all structures, indicates that the β-turn structure is a characteristic for all conformers, including the global minimum energy structure. The superpositioning of the average local minimum energy structure $\overline{Compstatin}^{Local}$ and the global minimum energy structure is given in Figure 1.

Global Optimization with TAD
The modified constrained global optimization was also applied to the Compstatin structure prediction problem. Initially TAD is used in combination with simple van der Waals overlap restraints to drive the distance violations to zero.

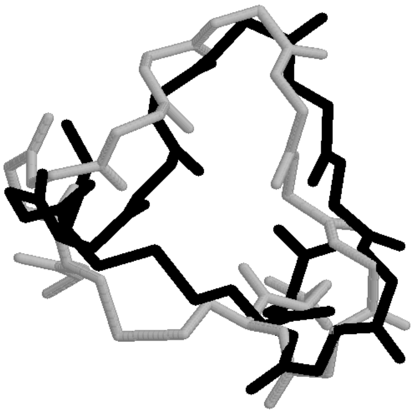

Figure 1: Superposition of global minimum (in black) and $\overline{Compstatin}^{Local}$ (in light grey) structures.

To gauge the performance of the combined αBB and TAD constrained approach, a comparison was made to an independent TAD method (DYANA [10]) for solving distance restraint problems. The coupled simulated annealing / TAD protocol was applied to a starting sample of 1000 randomly generated structures, from which a subset consisting of 20 conformers exhibiting the best target values were used to initialize a second set of runs. The five conformations with the best target function values were selected for further analysis, including initialization for constrained local minimizations with $E_{\mathrm{ECEPP/3}}$. The DYANA conformers satisfy the corresponding constraint, although their energy values are more than 70 kcal/mol higher than that of the global minimum energy structure. An analysis of the structural characteristics also indicates that the type I β-turn does not form along the Gln^5-Gly^8 b backbone for these structures. These results reflect the potential deficiencies of the independent TAD algorithm; that is, the simplified force field term is insufficient for sparse sets of distance restraints.

The use of TAD in the context of the global optimization approach surmounts this difficulty by using an iterative TAD scheme with two forms of the target function. The first set of TAD runs focuses on the reduction of the distance violations, while employing a simplified forcefield in the form of additional distance restraints to avoid atomic overlaps. This approach mimics the effects of a typical TAD approach for structure prediction. To ensure that these conformers provide low energy, this step is then followed by unconstrained minimization with a hybrid distance and ECEPP/3 energy objective function. Figure 2 shows a typical sequence for both the ECEPP/3 and distance violations energy during one solution of the upper bounding problem for Compstatin.

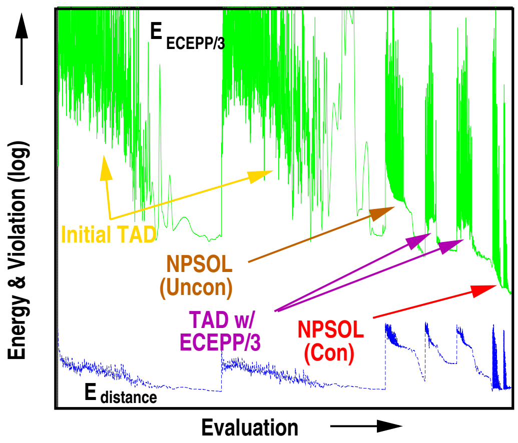

Figure 2: Log plot of $E_{\mathrm{ECEPP/3}}$ and $E^{distance}$ during a typical solution to the upper bounding problem for C3.

The results of the combined constrained global optimization and TAD algorithm can be assessed by examining the sequence of ECEPP/3 energies obtained from the solution of the upper bounding problems.When compared to the original algorithm, the TAD implementation augments the number of feasible starting points by more than a factor of two. This enhancement leads to earlier identification of low energy conformers. In particular, conformers with energies less than -70 kcal/mol, and thus lower in energy than the locally minimized PDB structures, are identified within the first 10 iterations of the global optimization approach. This property has important algorithmic implications, including the ability to fathom regions based on the current estimate of the global minimum.

Conclusions

Both experimental and theoretical methods exist for the prediction of protein structures. In both cases, additional restraints on the molecular system can be derived and used to formulate a nonconvex optimization problem. In this work, the traditional unconstrained problem was recast as a constrained global optimization problem, and applied to protein structure prediction using NMR data. Both the formulation and solution approach of this method differ from traditional techniques, which generally rely on the optimization of penalty-type target function using SA/MD protocols.

The approach was applied to the Compstatin structure prediction problem using both the original TAD approach and the coupled αBB-TAD approach. When considering basic structural features, such as the formation of a type I β-turn, the predicted structure was found to agree with results based on X-PLOR [6]. However, constrained global optimization was able to identify conformers with significantly lower energies than those obtained from either local minimization or independent TAD algorithms. In particular, the coupled αBB-TAD implementation consistently produced dense populations of low energy conformers.

Acknowledgments

The authors gratefully acknowledge financial support from the National Science Foundation, Air Force Office of Scientific Research, and the National Institutes of Health (R01 GM52032).

References

[1] Adjiman C.S., Androulakis I.P., and Floudas C.A., 1998a, A global optimization method, αBB, for general twice–differentiable NLPs – II. Implementation and computational results. *Comput. Chem. Eng.* **22**, 1159–1179.

[2] Adjiman C.S., Dallwig S., Floudas C.A., and Neumaier A., 1998b, A global optimization method, αBB, for general twice–differentiable NLPs – I. Theoretical advances. *Comput. Chem. Eng.* **22**, 1137–1158.

[3] Androulakis I.P., Maranas C.D., and Floudas C.A., 1995, αbb : A global optimization method for general constrained nonconvex problems. *J. Glob. Opt.* **7**, 337–363.

[4] Androulakis I.P., Maranas C.D., and Floudas C.A., 1997, Global minimum potential energy conformation of oligopeptides. *J. Glob. Opt.* **11**, 1–34.

[5] Berendsen H.J.C., Postma J.P.M., van Gunsteren W.F., DiNola A., and Haak J.R., 1984, Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690.

[6] Brünger A., 1992, X-PLOR, version 3.1 a system for x-ray crystallography and nmr. Yale University Press, New Haven, USA.

[7] Caves L.S.D., Evanseck J.D., and Karplus M., 1998, Locally accessible conformations of proteins: Multiple molecular dynamics simulations of crambin. *Protein Sci.* **7**, 649–666.

[8] Duan Y. and Kollman P.A., 1998, Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282**, 740–744.

[9] Floudas C.A., 2000, *Deterministic Global Optimization : Theory, Methods and Applications*. Nonconvex Optimization and its Applications, Kluwer Academic Publishers.

[10] Güntert P., Mumenthaler C., and Wüthrich K., 1997, Torsion angle dynamics for nmr structure calculation with the new program dyana. *J. Mol. Biol.* **273**, 283–298.

[11] Jain A., Vaidehi N., and Rodriguez G., 1993, A fast recursive algorithm for molecular dynamics simulation. *J. Comp. Phys.* **106**, 258–268.

[12] Klepeis J.L., Androulakis I.P., Ierapetritou M.G., and Floudas C.A., 1998, Predicting solvated peptide conformations via global minimization of energetic atom–to–atom interactions. *Comput. Chem. Eng.* **22**, 765–788.

[13] Klepeis J.L. and Floudas C.A., 1999, A comparative study of global minimum energy conformations of solvated peptides. *J. Comp. Chem.* (in press).

[14] Klepeis J.L., Floudas C.A., Morikis D., and Lambris J.D., 1999, Predicting peptide structures using nmr data and deterministic global optimization. *J Comp Chem* **20**, 1354–1370.

[15] Maranas C.D., Androulakis I.P., and Floudas C.A., 1996, A deterministic global optimization approach for the protein folding problem. In *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol. 23, (p. 133), American Mathematical Society.

[16] Morikis D., Assa-Munt N., Sahu A., and Lambris J.D., 1998, Solution structure of compstatin, a potent complement inhibitor. *Protein Sci.* **7**, 619–627.

[17] Némethy G., Gibson K.D., Palmer K.A., Yoon C.N., Paterlini G., Zagari A., Rumsey S., and Scheraga H.A., 1992, Energy parameters in polypeptides. 10. *J. Phys. Chem.* **96**, 6472.

[18] Rice L.M. and Brünger A.T., 1994, Torsion angle dynamics: Reduced variable conformational sampling enhances crystallographic structure refinement. *Proteins* **19**, 277–290.

[19] Sahu A., Kay B., and Lambris J., 1996, Inhibition of human complement by a c3-binding peptide isolated from a phage-displayed random peptide library. *J. Immunol.* **157**, 884–891.

[20] Standley D.M., Eyrich V.A., Felts A.K., Friesner R.A., and McDermott A.E., 1999, A branch and bound algorithm for protein structure refinement from sparse nmr data sets. *J Mol Bio* **285**, 1691–1710.