Original Article

Computational Study of Hydrolethalus Syndrome: D211G Mutation Destabilizes the Structural Stability of HYLS1 Leading to the Formation of Nuclear Inclusions

Paul M. Morgan[1], Yun-Ju Chen[2]

Objectives: To computationally determine the molecular structure of Hydrolethalus Syndrome 1 protein and the potential effects of D211G mutation on structural stability of the protein.
Methods: The structure of HYLS1 was predicted using two reputable algorithm servers: Robetta and QUARK. Robetta employs a combination of comparative modeling and *de novo* structure prediction methods, while QUARK is a template-free protein prediction server relying only on *de novo* structure prediction methods.
Results: Robetta predicted that 61% of the three dimensional structure of HYLS1 is intrinsically disordered, while 36% has well-defined secondary structural elements. The remaining 3% comprises a region of low complexity. HYLS1 is predicted to have 11 alpha helices and 2 beta strands.
Conclusions: Helix 9 of HYLS1 is stabilized by inter-residue electrostatic interaction and hydrogen bonding between Asp 211 and Arg 215. The stability is significantly disrupted by a D211G point mutation which causes Hydrolethalus Syndrome. This mutation may elicit random coil transition of an alpha helix, leading to the exposure of hydrophobic regions of the protein that aggregate to form nuclear inclusion bodies.

Key words: missense mutation, force fields, helix-to-coil transition, inclusion bodies, intrinsic disorder

Introduction

Hydrolethalus syndrome (HLS) is a lethal malformation syndrome that was discovered in Finland during a 1981 nationwide study on Meckel syndrome. Both of these syndromes are primarily identified by severe central nervous system malformation and polydactyly.[1-3] The incidence of HLS in the Finn-

From the [1]School of Medicine for International Students, and [2]Department of Medical Research, E-Da Hospital, I-Shou University, Kaohsiung, Taiwan.
Received: July 27, 2017 Accepted: November 16, 2017
Address reprint request and correspondence to: Paul M. Morgan, School of Medicine for International Students, I-Shou University, No. 8, Yida Road, Jiaosu Village, Yanchao District, Kaohsiung City 82445, Taiwan.
Tel: +886-966543608, E-mail: buddymorgan@gmail.com; isu10557002A@cloud.isu.edu.tw

ish population is 1 : 20,000 and only few cases have been reported outside of Finland.[1,3-4]

In HLS the cerebral hemispheres lie separated from each other at the base of the skull. As a consequence, the lateral ventricles open medially into the fluid-filled space between and on the top of the cerebral hemispheres.[5] The upper midline structures, corpus callosum and septum pellucidum, are absent.[6] There is also a midline deformity in the occipital bone dorsal to the foramen magnum giving rise to a keyhole-shaped defect. A Hypoplastic mandible is commonly observed. Short extremities, polydytaly and club foot are also common findings.[6,7]

Hydrolethalus syndrome in humans is caused by a missense mutation in the Hydrolethalus Syndrome 1 gene (HYLS1) located on chromosome 11q24.2.[1-3] This mutation results in an amino acid substitution of aspartic acid 211 by glycine in a well-conserved region of the protein with a function that is so far unidentified.[1,3] HYLS1 is predicted to encode a 299 amino acid polypeptide with a deduced molecular weight of 34.4 kDa. The Asp 211 is strictly conserved across various species from *C. elegans* to humans, suggesting that it is a critical amino acid for the normal function of the protein.[1-3]

Herein, I computationally explore the structure of HYLS1 at atomic level using *ab initio de novo* protein modeling. I also propose an explanation for the formation of inclusion bodies as a consequence of the A to G point mutation which perturbs the structural stability of the protein.

Materials and Methods

Ab Initio Folding

Predicting protein 3D structures from the amino acid sequence remains one of the major unsolved problems in computational biology.[8,9] This problem has baffled the scientific community for decades and remains one of the top 125 outstanding issues in modern science.[10]

Despite notable successes, we still have very limited ability to fold proteins by *ab initio* approaches, that is, to predict 3D structures of protein sequences without using template structures from other experimentally solved proteins.[8]

The difficulty in *ab initio* protein structure prediction is twofold. The first is the lack of decent force fields to accurately describe the atomic interactions which can be used to guide the protein folding simulations.[8] The second lies in the challenge to efficiently identify the global energy minimum which is supposed to be the native state of the protein - governed by the laws of thermodynamics.[8]

The structure of HYLS1 was predicted using two renowned algorithms: Robetta, developed by the University of Washington, and QUARK, developed by the University of Michigan Medical School. The Robetta server parsed the 299 amino acid sequence of HYLS1 into putative domains and structural models were generated after 720 hours of computing time.[9] Robetta employed a combination of comparative modeling and *de novo* structure prediction methods.[9]

QUARK is a template-free protein prediction server and as such relies only on *de novo* structure prediction methods.[8] The QUARK algorithm assembles full length protein models from fragments using replica-exchange Monte Carlo simulations which are guided by composite knowledge-based force fields.[8] QUARK, however, has a maximum capacity of 200 amino acids in any given protein. To overcome this challenge, HYLS1 was parsed into four distinct domains.[9] Only the fourth domain was submitted for modeling. Domain four is of great interest because it contains D211G mutation in a highly conserved region of the protein. QUARK generated several structural models for the wild-type and mutant form of

HYLS1 after 96 hours of computing time.

CASP

CASP (Critical Assessment of Methods of Protein Structure Prediction) biannual competition provides an independent mechanism for the assessment of current methods in protein structure modeling.[11] Biannually, from April through July, each year, structures about to be solved by crystallography or NMR are identified, and their sequences are made available to predictors.

From July to December, during the period of competition, experimental coordinates become available and tens of thousands of models are submitted by approximately 200 research groups worldwide.[8,11] The main goal of the CASP experiments is to obtain an in-depth and objective assessment of current potentialities in the area of protein structure prediction.

Robetta was the top-performing Homology server at CASP 2016 (Template-Based Modeling) and is one the most cited servers in Continuous Model Evaluation testing.[11] ITASSER-QUARK was ranked as the number one server in the world for template-free modeling in CASP9 2012 and CASP10 2014 experiments.[11]

Energy Force Fields

Energy force fields are mathematical functions used to accurately describe all atomic interactions in a protein during protein folding simulations (Table 1). The total energy function of the QUARK force field was derived from the sum of the eleven terms (Equation 1).[8]

All the energy terms are highly accurate and were derived from extensive statistical data of experimentally determined protein structures currently present in the Protein Data Bank.[8]

The models with the highest C-score from both servers were selected for further analysis and the PBD files were analyzed using Schrödinger Maestro 10.[12] The multiple sequence alignment and bioinformatics analysis was performed using jalview.

$$E_{\text{tot}} = E_{\text{prm}} + w_1 E_{\text{prs}} + w_2 E_{\text{ev}} + w_3 E_{\text{hb}} + w_4 E_{\text{sa}} + w_5 E_{\text{dh}} + w_6 E_{\text{dp}} + w_7 E_{\text{rg}} + w_8 E_{\text{bab}} + w_9 E_{\text{hp}} + w_{10} E_{\text{bp}}$$

Equation. 1 The eleven energy terms were segregated into three levels with reference to protein structural hierarchy. Level one, the atomic-level terms (Eprm, Eprs, and Eev); Level two, the residue-level terms (Ehb, Esa, Edh, and Edp), and; Level three, the topology-level terms (Erg, Ebab, Ehp, and Ebp).[8] The variables w1 5 0.10, w2 5 0.03, w3 5 0.03, w4 5 4.00, w5 5 0.40, w6 5 0.60, w7 5 1.00, w8 5 1.00, w9 5 0.05, and w10 5 0.10 were used as weighting factors to balance the energy terms.[8]

Table 1. Depiction of the mathematical functions for each of the eleven energy terms that govern atomic interactions. All eleven mathematical functions were integrated into the QUARK and ROBETTA algorithmic servers during the de novo protein folding simulation and prediction studies of HYLS1.[8,9]

	Energy Term	Energy Function
1	Backbone atomic pair-wise potential[8]	$E_{\mathrm{prm}}(i,j,r_{ij}) = -RT\log\left(\frac{N_{\mathrm{obs}}(i,j,r_{ij})}{r_{ij}^{\alpha}N_{\mathrm{obs}}(i,j,r_{\mathrm{cut}})}\right)$
2	Side-chain center pair-wise potentials[8]	$E_{\mathrm{prs}}(i,j,r_{ij}) = -RT\log\left(\frac{N'_{\mathrm{obs}}(i,j,r_{ij})}{r_{ij}^{\alpha'}N'_{\mathrm{obs}}(i,j,r_{\mathrm{cut}})}\right)$
3	Excluded volume[8]	$E_{\mathrm{ev}}(i,j,r_{ij}) = \begin{cases} (vdw(i)+vdw(j))^2 - r_{ij}^2 & \text{if } r_{ij} < vdw(i)+vdw(j) \\ 0 & \text{else} \end{cases}$
4	Hydrogen bonding[8]	$E_{\mathrm{hb}}(i,j,T_k) = \sum_{l=1}^{n_k}\frac{(f_l(i,j)-\mu_{kl})^2}{2\delta_{kl}^2}, \quad n_k = \begin{cases} 4 & k=1,2 \\ 3 & k=3,4 \end{cases}$
5	Solvent accessibility[8]	$E_{\mathrm{sa}} = \sum_{i=1}^{L}\lvert s_i - s_i^E\rvert$ $s_i = 1 - w\sum_{d(G_i,G_j)<9\text{Å}}\frac{A_{aa(j)}}{d^2(G_i,G_j)}$
6	Backbone torsion potential[8]	$E_{\mathrm{dh}} = -\sum_{i=2}^{L-1}\log(P(\phi_i,\psi_i\lvert aa(i),ss(i)))$
7	Fragment-based distance profile[8]	$E_{\mathrm{dp}} = -\sum_{(i,j)\in S_{\mathrm{dp}}}\log(N_{i,j}(d_{ij}))$
8	Radius of gyration[8]	$E_{\mathrm{rg}} = \begin{cases} 0 & r_{\min}\le r\le r_{\max} \\ (r_{\min}-r)^2 & r < r_{\min} \\ (r-r_{\max})^2 & r > r_{\max} \end{cases}$
9	Strand–helix–strand packing[8]	$E_{\mathrm{bab}} = \begin{cases} E_{\mathrm{pen}} & \text{left}-\text{handed} \\ 0 & \text{else} \end{cases}$
10	Helix packing[8]	$E_{\mathrm{hp}}(i,j) = -\log(P(d_{ij},\ \phi_{ij}))$
11	Strand packing[8]	$E_{\mathrm{bp}}(i,j) = -\log(P(aa(i),aa(j),T_{ij}))$

Results

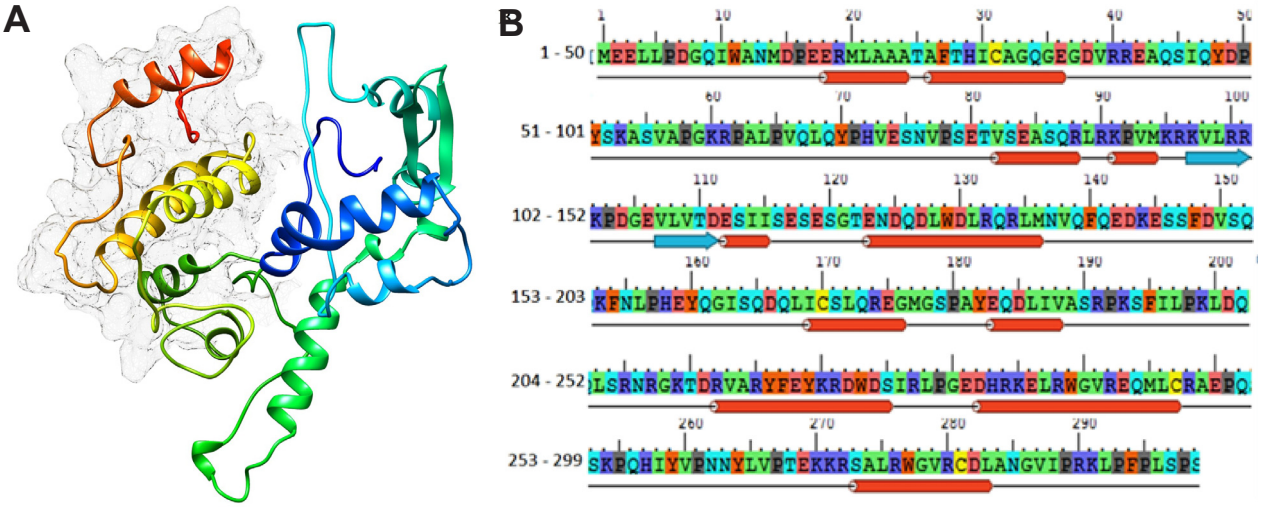

Fig. 1 (A) 3D structure prediction of HYLS1. Robetta predicts that 61% of the 3D structure of HYLS1 is intrinsically disordered, while 36% has well-defined secondary structural elements, primarily alpha helices. The remaining 3% comprises a region of low complexity. (B) Layout of the secondary structural elements of HYLS1. Robetta predicts that HYLS1 is composed of 11 alpha helices and 2 beta strands. The putative domain 4 highlighted with a transparent mesh surface (A) at the C-terminus (202 - 282) is strictly conserved across many species.

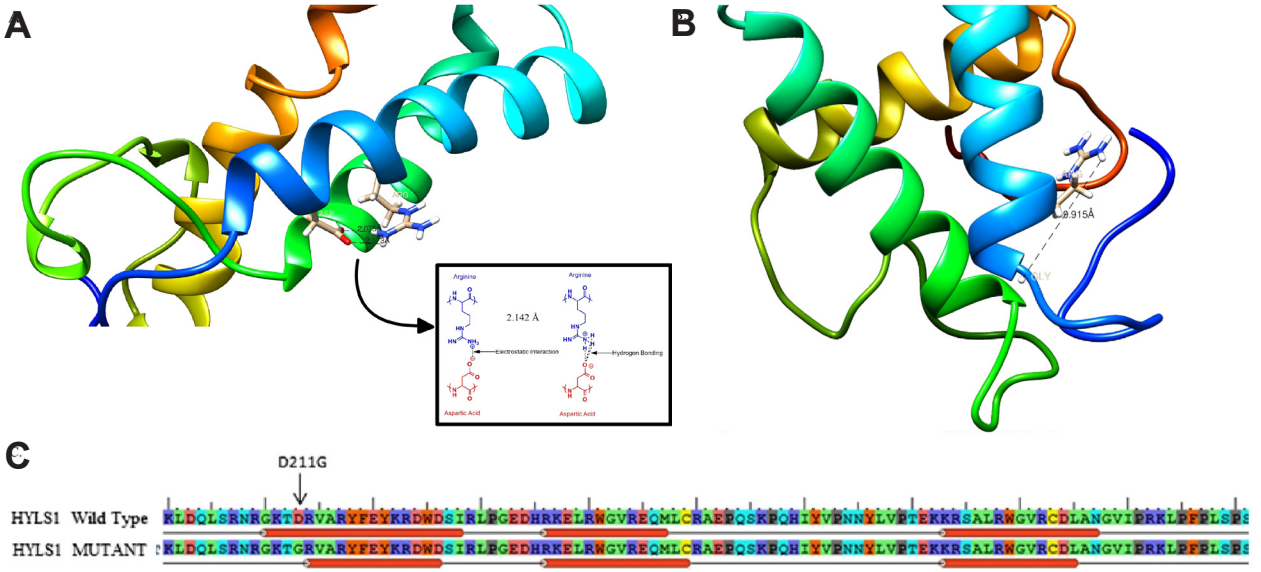

Fig. 2 (A) QUARK's atomic level resolution model of wild-type HYLS1 domain 4 (202 - 282) near the C-terminus. Helix 9 appears to be stabilized by inter-residue electrostatic interaction and hydrogen bonding between Asp 211 and Arg 215 (i + 4 residue). The average distance of 2.142 Å is well in agreement with that in reported studies. In the mutant form ((B) and (C)) the non-conservative point mutation Asp211Gly appears to significantly perturb helix formation and stabilization, thereby disrupting the electrostatic interaction and hydrogen bonding. Besides, glycine 211, being flexible and unable to form any favorable interactions with Arginine 215, may account for an average inter-residue distance of 9.915 Å that supports the helix disruption postulate.

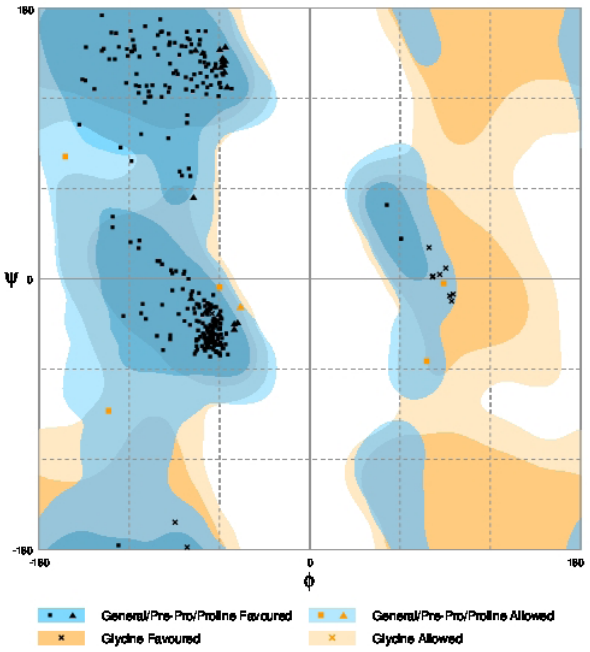

Fig. 3 Predicted Ramachandran plot for HYLS1 suggesting intrinsic disorder in a significant portion of HYLS1.[13]

Discussion

Etiology of HLS

The etiology of HLS is a missense mutation in the HYLS1 gene located on chromosome 11q24.2.[14] HLS has no known cure.[14] Recent advances in genome editing technologies have substantially improved our ability to make precise changes in the genomes of eukaryotic cells. CRISPR-Cas9 is emerging as a powerful tool in correcting genetic mutations which cause genetic diseases such as Down syndrome, spina bifida, anencephaly, and hydrolethalus syndrome.[15,16] CRISPRCas9, in particular, has the potential to treat the millions of patients who are impacted by a number of very serious diseases.[15] However, a cure for hydrolethalus syndrome through CRISPR-Cas9 is yet to be achieved.

HYLS1 Structure and Function

The function of the HYLS1 protein is still largely unknown and it lacks any known functional domains except a low-complexity region.[4] In addition, the protein is not homologous with any known protein family.[1,4] HYLS1 has orthologs among other species with a conserved polypeptide domain where the mutation site is located.[4]

The factors that drive the robustness and evolvability of proteins are still largely unknown.[17] Abrusan et al. suggests that different secondary structural elements of proteins (helices and strands) differ in their ability to tolerate mutations, and further demonstrates that it is caused by differences in the number of non-covalent residue interactions within these secondary structural units.[17,18] Based on those observations, human missense mutations that perturb secondary structure are more likely to be pathogenic than those that do not. The results of my computational study are in direct agreement.[14,17]

Aspartic acid is bulky and negatively charged, whereas glycine is the smallest amino acid with a neutral charge, facilitating turns in the polypeptide backbone.[19]

It is a well-studied phenomenon that certain amino acids are found more frequently in alpha helices (alanine, leucine, glutamic acid) while others are much less frequent (proline, glycine, aspartic acid).[20] Alanine has been shown to have the highest helix-forming propensity and glycine, the lowest.[20] This strongly supports the findings that the Asp211Gly mutation in HYLS1 significantly disrupts the formation and stabilization of helix 9. The destabilization of helix9 may trigger random coil transition of an alpha helix. Glycine, with no side chain, is a helix-breaker because backbone rotation is so unconstrained.[20,21]

Alpha helices are primarily stabilized by a favorable enthalpic contribution of 1 kcal/mol per residue from the formation of the backbone hydrogen bonds, van der Waals forces and side chain interactions that are more favorable in

the helix than the coil.[20]

The randomly coiled state is greatly favored by conformational entropy.[17,21] The entropic cost of fixing the backbone dihedral angles in forming an alpha helix is in the range of 1.5 – 2 kcal/mol at 25°C.[20] If the alpha helix is composed of alanine, the favorable interactions prevail, and the helix is more stable than the coil, but if the alpha helix is composed of glycine, the unfavorable interactions prevail and the coil is more stable than the helix.[21]

Missense Mutations, Inclusion Bodies, and Genetic Diseases

A single mutation in the amino acid sequence of a protein may also change the interactions between the side chains that affect the folding and stability of the protein. This can lead to the exposure of hydrophobic regions of the protein that aggregate with the same misfolded protein, or even a different protein, leading to the formation of inclusion bodies.[20] Mutations which cause diseases are significantly more destabilizing (i.e., having a larger effect on free energy of folding) than neutral mutations.[17] Mee et al. observed that wild-type HYLS1-transfected cells showed characteristically intense staining in the cytoplasm, whereas the mutant polypeptides are localized to nuclear inclusions.[1,14]

Inclusion bodies are often hallmarks of genetic diseases such as Parkinson's disease and front temporal dementia.[14] Interestingly, HYLS1 expression can be observed in multiple tissues being particularly high in the developing central nervous system (CNS).[14] Expression in the nervous system is evident in the spinal cord and in the dorsal root ganglia.[14] In the cephalic region of mouse embryo, HYLS1 is detected in the telencephalon, the midbrain, the medulla, the choroid plexus and the ganglionic eminence.[14]

When comparing the isoelectric point (pI value) with Kozlowski IPC tool, wild-type HYLS1 is predicted to have a pI value of 6.69 versus 7.1 for the mutant protein.[22] The pI is the pH at which a protein carries no net electrical charge in solution. As a consequence, the solubility of the protein is significantly diminished, resulting in precipitation and formation of inclusion bodies (i.e., cytoplasmic and/or nuclear) in the cell. The results suggest that HYLS1 D211G mutant is significantly less soluble at physiological pH (7.36) as compared to its wild-type counterpart.[22,23] Additionally, wild-type HYLS1 carries a net charge of -1.9, while HYLS1 mutant carries a net charge of -0.9 in the cytoplasm.[18]

HYLS1 Intrinsic Disorder

Bioinformatics studies primarily based on the Pfam database suggests that HYLS1 has a region of low complexity between amino acids 112 and 120.[14] Such regions are typical for structurally less complex proteins. The conventional protein structure-function concept suggests that an amino acid sequence specifies a primarily fixed three-dimensional structure that is a prerequisite to protein function. In contrast to the dominant view, many proteins are emerging today as functional proteins in the disordered state.[23,24]

For ordered proteins, the ensemble members all have the same time-averaged canonical set of Ramachandran angles along their backbones. Nevertheless, for intrinsically disordered proteins, the ensemble members have different, typically dynamic, Ramachandran angles.[13,23,24]

Intrinsically disordered proteins are therefore difficult to characterize by current conventional biophysical techniques (i.e., x-ray crystallography, NMR spectroscopy, circular dichroism spectroscopy).[13,24] As such, computational methods may currently be the most practical method and perhaps the only option for exploring the molecular structure of HYLS1 with atomic level resolution.

HLS and Computational Medicine

In conclusion, many studies have put forth evidence that HYLS1 may be a transcription factor involved in the developmental process.[1-3] Nevertheless, there are still many unanswered questions about HYLS1 that may require computational and laboratory explorations.

For instance, can we validate the computationally predicted 3D structure with X-ray crystallography and NRM spectroscopy in the future? Will technology bring forth new methods for studying the structure of intrinsically disordered proteins? How many partners in the human proteome does HYLS1 directly interact with? How does the Asp211Gly mutation perturb these physical interactions to bring forth Hydrolethalus Syndrome?

In a subsequent study, I hope to use computational medicine and molecular dynamics to develop a quantitative computer model in an attempt to explain how the Asp211Gly mutation disrupts the global structure of HYLS1 leading to Hydrolethalus Syndrome.

Acknowledgements

The author would like to thank I-Shou University School of Medicine for International Students, Republic of China, Taiwan, for the opportunity to continue my research career as a scientist in the middle of a busy physician training course. The author also thanks Paul Morgan Sr. for editing the manuscript.

Disclosure

The author declares no conflict of interest concerning the materials and methods used in this study or the findings specified in this manuscript.

References

1. Dammermann A, Pemble H, Mitchell BJ, et al: The hydrolethalus syndrome protein HYLS-1 links core centriole structure to cilia formation. Genes Dev 2009;23:2046-59.
2. Paetau A, Honkala H, Salonen R, et al: Hydrolethalus syndrome: neuropathology of 21 cases confirmed by HYLS1 gene mutation analysis. J Neuropathol Exp Neurol 2008;67:750-62.
3. Visapää I, Salonen R, Varilo T, et al: Assignment of the locus for hydrolethalus syndrome to a highly restricted region on 11q23-25. Am J Hum Genet 1999;65:1086-95.
4. Honkala H, Lahtela J, Fox H, et al: Unraveling the disease pathogenesis behind lethal hydrolethalus syndrome revealed multiple changes in molecular and cellular level. Pathogenetics 2009;2:2.
5. Ämmälä P, Salonen R: First-trimester diagnosis of hydrolethalus syndrome. Ultrasound Obstet Gynecol 1995;5:60-2.
6. Salonen R, Herva R: Hydrolethalus syndrome: J Med Genet 1990;27:756-9.
7. Salonen R, Herva R, Norio R: The hydrolethalus syndrome: delineation of a "new", lethal malformation syndrome based on 28 patients. Clin Genet 1981;19:321-30.
8. Xu D, Zhang Y: Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 2012;80:1715-35.
9. Kim DE, Chivian D, Baker D: Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 2004;32:W526-31.
10. Dill KA, Ozkan SB, Weikl TR, et al: The protein folding problem: when will it be solved? Curr Opin Struct Biol 2007;17:342-6.
11. Moult J, Fidelis K, Kryshtafovych A, et al: Critical assessment of methods of protein structure prediction (CASP) — round x. Proteins 2014;82 Suppl 2:1-6.
12. Shaw DE. Desmond Molecular Dynamics System. version 3.1, Maestro-Desmond Interoperability Tools; Schrödinger.
13. Lovell SC, Davis IW, Arendall WB 3rd, et al: Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins 2003;50:437-50.
14. Mee L, Honkala H, Kopra O, et al: Hydrolethalus syndrome is caused by a missense mutation in a novel gene HYLS1. Hum Mol Genet 2005;14:1475-88.
15. Lundberg AS, Novak R: CRISPR-Cas gene editing to cure serious diseases: treat the patient, not the germ line. Am J Bioeth 2015;15:38-40.
16. Mentis AF: Epigenomic engineering for Down syndrome. Neurosci Biobehav Rev 2016;71:323-27.
17. Abrusán G, Marsh JA: Alpha helices are more robust to mutations than beta strands. PLOS Comput Biol 2016;12:e1005242.
18. Roberts CJ: Non-native protein aggregation kinetics. Biotechnol Bioeng 2007;98:927-38.
19. Holtzer ME, Holtzer A: Alpha-helix to random coil transitions: determination of peptide concentration from the CD at the isodichroic point. Biopolymers 1992;32:1675-77.

20. Pace CN, Scholtz JM: A helix propensity scale based on experimental studies of peptides and proteins. Biophys J 1998;75:422-7.
21. Baker EG, Bartlett GJ, Crump MP, et al: Local and macroscopic electrostatic interactions in single α-helices. Nat Chem Biol 2015;11:221-8.
22. Tyedmers J, Mogk A, Bukau B: Cellular strategies for controlling protein aggregation. Nat Rev Mol Cell Biol 2010;11:777-88.
23. Wright PE, Dyson HJ: Intrinsically disordered proteins in cellular signaling and regulation. Nat Rev Mol Cell Biol 2015;16:18-29.
24. Dunker AK, Lawson JD, Brown CJ, et al: Intrinsically disordered protein. J Mol Graph Model 2001;19:26-59.
25. Roy A, Kucukural A, Zhang Y: I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 2010;5:725-38.
26. Marchler-Bauer A, Bo Y, Han L, et al: CDD/SPARCLE functional classification of proteins via subfamily domain architectures. Nucleic Acids Res 2017;45:D200-3.
27. Marchler-Bauer A, Derbyshire MK, gonzales NR, et al: CDD: NCBI's conserved domain database. Nucleic Acids Res 2015;43:D222-6.
28. Marchler-Bauer A, Lu S, Anderson JB, et al: CDD: a conserved domain database for the functional annotation of proteins. Nucleic Acids Res 2011;39:D225-29.
29. Buchan DW, Minneci F, Nugent, et al: Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res 2013;41:W340-48.
30. Yang J, Yan R, Roy A, et al: The I-TASSER Suite: protein structure and function prediction. Nat Methods 2015;12:7-8.
31. Pettersen EF, Goddard, TD, Huang, CC, et al: UCSF Chimera - a visualization system for Eeploratory research and analysis: J Comput Chem 2004;25:1605-12.
32. Marchler-Bauer A, Bryant SH: CD-Search: protein domain annotations on the fly. Nucleic Acids Res 2004;32:W327-31.
33. Zhang Y: I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 2008;9: 40.
34. Abascal F, Valencia A: Automatic annotation of protein function based on family identification. Proteins 2003;53:683-92.
35. Altschul SF, Madden TL, Schaffer AA, et al: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389-402.
36. Bateman A, Coin L, Durbin R, et al: The Pfam protein families database. Nucleic Acids Res 2004;32:D138-41.
37. Salsbury FR Jr: Molecular Dynamics Simulations of protein dynamics and their relevance to drug discovery. Curr Opin Pharmacol 2010;10:738-44.