Analysis of the Three Dimensional Structure of the CXGXC Motif in the CMGCC and CAGYC Regions of α- and β-Subunits of Human Chorionic Gonadotropin: Importance of Glycine Residue (G) in the Motif

KENGO KINOSHITA**, MASAMI KUSUNOKI AND KIYOSHI MIYAI***

Institute for Protein Research, Osaka University, 3-2 Yamadaoka Suita-shi, 565-0871 Osaka Japan
*Japan Science Corporation
**Present address: Institute of Medical Science, University of Tokyo, Structure and Function of Biomolecules, SORST, JST, Tokyo, Japan
***Osaka University (Professor Emeritus), Osaka Japan

Abstract. The three dimensional structures of the C1X2G3(X3)C4C5 motif of hCG, which is considered to be important for noncovalent assembly of the α- and β-subunits of glycoprotein hormones were analyzed to assess the importance of glycine (Gly) (G) at site X3 in the motif by the conformational energy calculation using computational procedures. In the C1M2G3(X3)C4C5 motif of the α-subunit, Ramachandran plot analysis showing the allowed area of the dihedral angles demonstrated that only a Gly residue was allowed at site X3. In calculating collision with surrounding atoms as a monomer the possible main chain models of the C1A2G3(X3)Y4C5 motif in the β-subunit showed that only alanine (Ala) (A) or Gly at site X3 is allowed to alleviate the collision with the cysteine (Cys) (C) residues which form a disulfide bridge. A mutant of the β-subunit with the C1A2A3(X3)Y4C5 motif (Ala at site X3) may not compose a heterodimer with the α-subunit because of interference of intermolecular hydrogen bond formation. These findings indicate that the Gly residue at site X3 (G) in the motif is essential for heterodimer formation of glycoprotein hormones. The significance of similar motifs found in various human proteins other than glycoprotein hormones was suggested.

Key words: CXGXC motif, Ramachandran plot analysis, hCG, Glycoprotein hormones, TSH deficiency

THE human glycoprotein hormones, hTSH, hLH, hFSH and hCG, are composed of two noncovalently linked α- and β-subunits (chains). The amino acid sequences of α-subunits (common α-subunit) are identical in all glycoprotein hormones, and those of β-subunits are specific for each hormone.

In 1971, Miyai et al. [1] reported the first case of congenital glycoprotein hormone (TSH) deficiency, and subsequent studies revealed that the disease is caused by a point mutation (G to A) in the hTSH β-subunit gene which changes a glycine (Gly) (G) to an arginine (Arg) (R) in the CAGYC region composed of cysteine (Cys) (C)-alanine (Ala) (A)-glycine-tyrosine (Tyr) (Y)-cysteine [2–5]. Recently a case of hypogonadism caused by a mutation in the hLH β-subunit gene which changes a Gly to an aspartic acid (Asp) (D) in the region was reported [6]. On the other hand, a similar amino acid sequence, cysteine-methionine (Met) (M)-glycine-cysteine-cysteine (CMGCC), is located in the common α-subunit. We performed experiments of site-directed mutagenesis in the CAGYC [3, 5, 7] and CMGCC [8, 9] regions of hTSH and hCG. We also analyzed the amino acid sequences of these regions in wild type glycoprotein hormones of various species and demonstrated that the regions are conserved among all of the known α- and β-subunits of glycoprotein hormones [5, 8]. The findings supported...
our hypothesis that the motif of CXGXC (X: any residue) may play an important role in biosynthesis, particularly in the noncovalent assembly of the subunits of glycoprotein hormones.

In 1994, Lapthorn et al. [10] and Wu et al. [11] reported the three dimensional structure of hCG and revealed that a common disulfide bridge (bond) \( (C_s) \) and a unique disulfide bridge, a cystine-knot \((C_k)\), are located in the regions \( (C_kMGCC_k, C_kAGYC_k) \), suggesting that two disulfide bridges at both ends are essential to the region.

In the present study, we analyzed the three dimensional structure of the regions by computational procedures to assess the importance of Gly \((G_3)\) for the formation of the motif of \( C_1X_2G_3(X_3)X_4C_5 \).

**Methods**

The computational procedures for conformational energy calculation are as follows. The crystal structure of hCG with a PDB id of 1HCN was used for the conformational energy calculation for two representative motifs of glycoprotein hormones, CMGCC (28–32) in the \( \alpha \)-chain and CAGYC (34–38) in the \( \beta \)-chain. Both motifs consist of five amino acid residues. All the Cys residues of both motifs are involved in disulfide bonds and some of them form the cystine knot structure, resulting in a rigid conformation by the disulfide bridges (Fig. 1). We performed the main chain conformational calculations for our models of five residues in which the positions of Cys residues at both ends were fixed in space, and bond lengths and bond angles of the peptides were kept at standard values with the loop closure algorithm [12]. Under these conditions all the possible conformations of the five residues segment are generated and counted. The number of conformational freedom is 8 for the main chain dihedral angles \( \phi \) and \( \psi \). Calculation of all the combinations of the 8 dihedral angles corresponds to exploring all the main conformations of the peptide. Among these main chain conformations, we selected only the conformations where the peptide chain atoms do not collide with surrounding atoms in a chain of hCG (PDB-ID 1HCN) and defined them here as the possible main chain conformations. In more detail, the \( \psi \) angle at the N-terminus and the \( \phi \) angle at the C-terminus were changed in 1 degree steps. For each combination of \( \phi \) and \( \psi \) angles the analytical solutions for the middle 6 main chain dihedral angles were obtained by the loop closure algorithm. In this

---

**Fig. 1.** Disulfide bonds in the \( \alpha \)-(a) and \( \beta \)-(b) chains of hCG. Thick lines indicate the polypeptide chain of hCG. Thin and dual lines both denote disulfide bonds and the latter refers to the ones involved in cystine knots. Rectangles and circles indicate \( \beta \)-strands \( \beta 1 \) through \( \beta 4 \) and cysteine residues, respectively. **N** and **C** denote the N- and C-terminii of the polypeptide.
way, all the possible main chain conformations were generated and counted.

In calculating collision with surrounding atoms as a dimer, 10418 and 6110 sets of possible main chain conformations are obtained in α- and β-chains, respectively.

**Results**

*Analysis of the middle Gly residue in the CMGCC motif of the α-subunit*

The conformational energy calculation for the representative motif C₁M₂G₃C₄C₅ was performed based on the crystal structure of hCG with a PDB id of 1HCN. All the Cys residues of the motif (C₁, C₄, C₅) are involved in disulfide bonds as shown in Fig. 1. In order to explore the plausibility of various amino acid residues at the middle three residue sites for the five residue motifs, we performed the main chain conformational calculation for the following model system described in the section of computational procedure. Under structural constraints imposed by the two Cys residues (C₁ and C₅) at both ends of the motif, we explored all the possible conformations to be assumed by the main chain of the motif and found the following results. Fig. 2 shows the Ramachandran diagram which illustrates the allowed areas of the dihedral angles (ϕ and ψ) for poly-L-alanine (Fig. 2c) and poly-glycine (Fig. 2d) [13]. The dihedral angles for amino acids other than Gly and proline (Pro) (P) in 12 proteins were found to be distributed mostly in the allowed areas for

---

**Fig. 2.** Ramachandran diagram. The shaded bars with scales in (a) and (b) denote number of sets of possible main chain conformations. (a) Ramachandran plot of residue 30 for the possible main chain conformations in the CMGCC region of the α-chain (b) Ramachandran plot of residue 36 for the possible main chain conformations in the CAGYC region of the β-chain (c) Limited (■) and ordinary (●) allowed areas for poly-L-alanine [13] (d) Limited (■) and ordinary (●) allowed areas for poly-glycine [13]
poly-L-alanine [14]. The regions of the Ramachandran plot analysis for the CMGCC motif (Fig. 2a) are located not in the allowed areas for poly-L-alanine (Fig. 2c) but only in those for poly-glycine (Fig. 2d). It indicates that the severe conformational restrictions at X₃ of the C₁M₂G₃(X₃)C₄C₅ motif allow only a Gly residue at this position (G₃). Fig. 3 illustrates the fixed averaged structure of the CMGCC motif in the α-chain showing the stable structure.

**Fig. 3.** The stick models of the peptide chains of the CMGCC motif in the α-chain (a) and the CAGYG motif in the β-chain (b) for the crystal structure. Carbon, nitrogen, oxygen, sulfur atoms are colored in gray, cyan, red, and yellow, respectively. The stick models in green and blue are computer generated averaged models and correspond to the conformations with the middle Gly having positive and negative φ values, respectively.

**Analysis of the middle Gly residue in the CAGYC motif of the β-subunit**

We calculated the conformation of the C₁A₂G₃Y₄C₅ motif based on the crystal structure of hCG in the same way as mentioned above. Again the two Cys residues (C₁ and C₅) at the end of the motif are involved in the disulfide bond. In Ramachandran diagram all regions for the CAGYC motif (Fig. 2b) are located in the allowed areas for poly-glycine (Fig. 2d) and a part of them are located in the area for poly-L-alanine (Fig. 2c) as well. Thus it indicates that the main chain conformational restrictions at site X₃ in the C₁A₂G₃(X₃)Y₄C₅ motif allow a Gly residue and also residues other than Gly at this position. Fig. 3 shows the diversity of the averaged structure of the CAGYC motif in which the structure with positive φ is different from that with negative φ.

Then we built an ensemble of peptide chain models for the motif which is allowed in the conformational space, and averaged them. We generated a Cα atom at X₃ in the C₁A₂G₃(X₃)Y₄C₅ motif as a standard conformer for the averaged peptide model. Fig. 4 shows that this atom is in a position to collide with the Cys residues (B9 C₉ (interatomic distance: 1.8 Å) and B57 S₃ (2.2 Å) in the β-chain) which are important for the cystine knot formation (Fig. 1b). It is possible to alle-
Table 1. The CXGXC motifs found in human proteins prepared using the database of 198 sequences matched to the pattern cxgxc in the iProClass with homo sapiens in organism obtained from the Protein Information Resource (PIR), June 16 EDT 2004 (http://pir.georgetown.edu/) [16] Refer to abbreviations

<table>
<thead>
<tr>
<th>CXGXC</th>
<th>Proteins*</th>
<th>CXGXC</th>
<th>Proteins*</th>
<th>CXGXC</th>
<th>Proteins*</th>
<th>CXGXC</th>
<th>Proteins*</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAGAC</td>
<td>1b, 1c, 2, 3c, 4</td>
<td>CEGLC</td>
<td>39, 40a</td>
<td>CLGKC</td>
<td>45b, 68</td>
<td>CRGNC</td>
<td>30b, 30f</td>
</tr>
<tr>
<td>CAGCC</td>
<td>5a, 5b, 5c</td>
<td>CEGPC</td>
<td>41a, 42a</td>
<td>CLGLC</td>
<td>69</td>
<td>CRGSC</td>
<td>91, 92a, 92b</td>
</tr>
<tr>
<td>CAGGC</td>
<td>6a, 6b, 6f</td>
<td>CEGSC</td>
<td>43, 44</td>
<td>CLGNC</td>
<td>70a</td>
<td>CSGAC</td>
<td>3b, 93</td>
</tr>
<tr>
<td>CAGNC</td>
<td>7</td>
<td>CFGKC</td>
<td>45a</td>
<td>CLGPC</td>
<td>1a, 70b</td>
<td>SGGCC</td>
<td>94b</td>
</tr>
<tr>
<td>CAGRC</td>
<td>8</td>
<td>CFGQC</td>
<td>19c, 46</td>
<td>CLGTC</td>
<td>30d</td>
<td>CSGHC</td>
<td>95, 96</td>
</tr>
<tr>
<td>CAGSC</td>
<td>9b, 9c, 9d, 9e, 9f, 9g, 10</td>
<td>CGGCC</td>
<td>47a, 47b</td>
<td>CLGVC</td>
<td>71</td>
<td>CSGIC</td>
<td>26a</td>
</tr>
<tr>
<td>CAGTC</td>
<td>11</td>
<td>CGGHC</td>
<td>50</td>
<td>CMGCC</td>
<td>72</td>
<td>GSGPC</td>
<td>1d, 99</td>
</tr>
<tr>
<td>CAGWC</td>
<td>12e</td>
<td>CGGMC</td>
<td>51</td>
<td>CMGMC</td>
<td>74</td>
<td>CSGRC</td>
<td>41b</td>
</tr>
<tr>
<td>CAGYC</td>
<td>12a, 12b, 12c, 12d, 12f, 12g, 13, 14, 15, 16, 17a</td>
<td>CGGDC</td>
<td>48a, 48b, 48c, 49</td>
<td>CGGMC</td>
<td>75</td>
<td>CSGVC</td>
<td>102</td>
</tr>
<tr>
<td>CAGTC</td>
<td>18</td>
<td>CGGRC</td>
<td>54</td>
<td>CGGMC</td>
<td>76a, 76b, 76c, 76d, 76e</td>
<td>CSGHC</td>
<td>17b</td>
</tr>
<tr>
<td>CAGCC</td>
<td>19a, 20, 21a</td>
<td>CHGDC</td>
<td>3d</td>
<td>CGGMC</td>
<td>77, 78</td>
<td>CTGCC</td>
<td>94a, 103</td>
</tr>
<tr>
<td>CAGKC</td>
<td>22</td>
<td>CHGEC</td>
<td>3a</td>
<td>CGGMC</td>
<td>79</td>
<td>CTGKC</td>
<td>76i, 104, 105</td>
</tr>
<tr>
<td>CAGLC</td>
<td>23, 24, 25</td>
<td>CHGGC</td>
<td>37c</td>
<td>CGGMC</td>
<td>80</td>
<td>CTGKC</td>
<td>106</td>
</tr>
<tr>
<td>CAGNC</td>
<td>26b</td>
<td>CHGC</td>
<td>57</td>
<td>CPGDC</td>
<td>81a, 81b</td>
<td>CTGTC</td>
<td>107</td>
</tr>
<tr>
<td>CAGRC</td>
<td>29, 30a</td>
<td>CCGAC</td>
<td>58</td>
<td>CPGEC</td>
<td>82</td>
<td>CVGHC</td>
<td>109</td>
</tr>
<tr>
<td>CAGVC</td>
<td>31</td>
<td>CCGMC</td>
<td>59</td>
<td>CGRGC</td>
<td>83</td>
<td>CVGHC</td>
<td>110</td>
</tr>
<tr>
<td>CDGEC</td>
<td>32</td>
<td>CCGQC</td>
<td>37b</td>
<td>CGRGC</td>
<td>84</td>
<td>CVGHC</td>
<td>111</td>
</tr>
<tr>
<td>CDGEC</td>
<td>3e, 3f</td>
<td>CGISC</td>
<td>59</td>
<td>CGRC</td>
<td>85</td>
<td>CVGHC</td>
<td>112</td>
</tr>
<tr>
<td>CDGGC</td>
<td>33, 34</td>
<td>CIGTC</td>
<td>60</td>
<td>CGQC</td>
<td>86</td>
<td>CVGHC</td>
<td>113</td>
</tr>
<tr>
<td>CDGEC</td>
<td>35, 36</td>
<td>CKGFC</td>
<td>61, 62a</td>
<td>CIGQC</td>
<td>87</td>
<td>CVGHC</td>
<td>114</td>
</tr>
<tr>
<td>CDGRC</td>
<td>7a</td>
<td>CGAC</td>
<td>36</td>
<td>CIGGC</td>
<td>88</td>
<td>CVGHC</td>
<td>115</td>
</tr>
<tr>
<td>CDGSC</td>
<td>38</td>
<td>CLGCC</td>
<td>66, 67</td>
<td>CIGTC</td>
<td>89</td>
<td>CVGHC</td>
<td>116</td>
</tr>
</tbody>
</table>

Abbreviations: A, Ala: alanine; C, Cys: cysteine; D, Asp: aspartic acid; E, Glu: glutamic acid; F, Phe: phenylalanine; G, Gly: glycine; H, His: histidine; I, Ile: isoleucine; K, Lys: lysine; L, Leu: leucine; M, Met: methionine; N, Asn: asparagine; P, Pro: proline; Q, Gln: glutamine; R, Arg: arginine; S, Ser: serine; T, Thr: threonine; V, Val: valine; W, Trp: tryptophan; Y, Tyr: tyrosine

viate this collision by modifying slightly the peptide conformation but the residues other than Gly and Ala would not be accommodated considering the Cβ direction.

Furthermore a mutant of G36A (C_1,C_2,A_3,Y_4,C_5) may be expressed as protein with a normal conformation, a heterodimer of hCG with physiological function would not be formed because an inter molecule hydrogen bond between Gly36 (G_3) of β-chain and serine (Ser) (S) 34 of α-chain which are important to dimer formation cannot be bridged (Fig. 4).

Discussion

In the present study, the necessity of the middle Gly residue (G_3) at site X_3 of the C_1,X_2,X_3,X_4,C_5 motif was demonstrated theoretically by conformational energy calculation. In the CMGCC motif of the α-chain only Gly was allowed to have the appropriate dihedral angles (φ and ψ) as shown by the Ramachandran plot analysis. In the CAGYC motif of the β-chain, only Gly or Ala at site X_3 is allowed to avoid collision between their side chain at Cβ and a cystine knot. The motif of C_1,A_2,Y_3,C_4 (Ala at X_3) may interfere with formation of the intermolecular hydrogen bond which is important for dimer formation of α- and β-chains.

These theoretical explanations are consistent with findings of previous studies as follows. Only Gly has been found at site X_3 in the motif of C_1,X_2,X_3,X_4,C_5 in various wild type α- and β-subunits [5, 8]. Site directed mutagenesis experiments have shown that alterations of this Gly to Ala, Asp or Arg in the CMGCC region of α-subunits [8, 9] and to Asp or Arg in the CAGYC region of β-subunits [3, 5, 7] resulted in loss...
of their biological and/or immunological activities of hTSH and hCG. Patients with TSH deficiency caused by a point mutation of the β-subunit gene that changed the Gly to Arg [2–5], and those with LH deficiency in which the Gly changed to Asp [6] have been reported.

Possible residues at sites X\textsubscript{2} and X\textsubscript{4} of the C\textsubscript{3}X\textsubscript{2}G\textsubscript{X}\textsubscript{C}\textsubscript{5} motif are considered as follows.

In α-subunits of wild type [5, 8] and mutant [8, 9] glycoprotein hormones, various amino acids (lysine (Lys) (K), methionine, arginine, serine, threonine (Thr) (T) and valine (Val) (V)) are found at site X\textsubscript{4}. The whole side chain of the X\textsubscript{4} site is exposed to the solvent and hence non-hydrophobic amino acid residues can occupy this position without losing the physiological function. Since Val is the most hydrophobic residue, residues at the X\textsubscript{4} site may be a non-hydrophobic residue with ΔG ≤ 1.5 kcal/mol for Val where ΔG is the free energy of transfer from organic solvent to water and an amino acid with the more positive values is the more hydrophobic [15]. But it is possible that even hydrophobic residues with its side chain tip having hydrophilic groups may occupy this site. In this case, another hydrophobic residue has to bury the hydrophobic part of the site chain at site X\textsubscript{4} near to the main chain. Met\textsubscript{41} of the β-chain is close to site X\textsubscript{4} of the α-chain and satisfies this requirement. According this reasoning, Tyr which is more hydrophobic but has a hydrophilic side chain tip, may be accepted at site X\textsubscript{2} of the α-chain. As for the X\textsubscript{4} site, the Cys 31 residue forms a disulfide bond with Cys 7 at the N-terminus of the α-chain (Fig. 1a) and serves to reduce the conformational freedom. This N-terminal region occupies a large surface area of the dimer interface between the α- and β-chains. Thus, the residue at X\textsubscript{4} to be Cys is functionally most important for the dimer formation. Replacement of the Cys residue with another amino acid is allowable for assuming the authentic conformation but may lead to denaturation of protein because of the failure of dimer formation. Thus the C\textsubscript{3}X\textsubscript{2}G\textsubscript{X}\textsubscript{C}\textsubscript{5} motif of the α-chain is described as C\textsubscript{π}GCC, where π is a non-hydrophobic residue as described above.

In the β-chain the residue preference at site X\textsubscript{2} (alanine, aspartic acid, glutamic acid, glycine, methionine, serine and threonine) [3, 5, 7, 8] is similar to the α-chain. Theoretical examination of the residue preference at site X\textsubscript{4} of the β-chain is difficult. The tip of the side chain at this site is exposed to solvent but the nearer part of the side chain to the main chain is buried. In the same way as described in the α-chain, a suitable residue of Arg 6 is located near site X\textsubscript{4} of the β-chain. Although Arg is hydrophilic, its side chain is long and the side chain other than the tip is hydrophobic. Considering the characteristic of this site, the allowable residues at site X\textsubscript{4} is widespread. We can exclude positively charged residues such as Lys, Arg for site X\textsubscript{4} because of unfavorable interactions. Thus the C\textsubscript{3}X\textsubscript{2}G\textsubscript{X}\textsubscript{C}\textsubscript{5} motif of the β-chain may be C\textsubscript{π}G[!KR]C, where !KR means the residues other than Lys and Arg.

A number of similar motifs of CXGXC have been found in various human proteins as shown in Table 1. Further studies are required to elucidate the significance of the CXGXC motif and the importance of Gly residue in the motif.

Acknowledgments

This work was supported in part by Research and Development for Applying Advanced Computational Science and Technology of Japan Science and Technology Corporation.

References


