

Annexe du chapitre 4 : Sélection des positions

Cette annexe détaille la méthode de sélection des positions actives employée dans le chapitre 4. Pour les tests avec une seule position active, comme les temps de calculs le permettent toutes les positions sont testées. Il y a huit cent quatre tests par algorithme. Pour les autres groupes de tests (cinq, dix, vingt et trente positions actives), cinq sélections sont effectuées pour chacune des neuf protéines, c'est-à-dire quarante-cinq tests par algorithme.

Pour définir complètement les tests, il faut décrire le choix des positions actives. Il y a peu d'intérêt à tester des situations dans lesquelles les positions actives n'interagissent pas ou faiblement entre elles. En effet, s'il existe une position active i dont chaque résidu est sans interaction avec tous les résidus possibles aux autres positions actives, déterminer le meilleur état pour i est proche du test dans lequel i est la seule position active de la sélection. Cela nous ramène vers les tests à une seule position active. Ainsi le choix des positions actives ne se fait pas par tirage aléatoire, car le risque d'obtenir des positions avec peu d'interactions est trop grand. Il se fait sous contraintes d'interaction. Pour cela, nous utilisons la notion de voisinage de proteus, introduite en 2.5.1. On dit qu'un ensemble de N positions \mathcal{V} est un « voisinage fort » pour le seuil s_{vois} si toutes les paires de positions (i et j) de \mathcal{V} sont voisines pour s_{vois} .

Pour sélectionner les positions actives, nous nous basons sur la liste des voisins de chaque position à un seuil donné. Le programme proteus peut fournir ces listes en mode verbeux, sans effectuer d'optimisation. Pour cela, nous utilisons le mode MONTECARLO avec une trajectoire de zéro pas. La recherche de voisinages forts se fait alors de la façon suivante :

1. s_{vois} est initialisé à 10.
2. construction des listes de voisins au seuil s_{vois}
3. recherche de cinq voisinages forts, par extension progressive d'une paire de positions voisines

4. si succès pour les neuf protéines arrêt, sinon diminution de s_{vois} de 1 kcal/mol
5. si $s_{vois} \geq 1$, retour à l'étape 2, sinon arrêt.

Nous obtenons les 45 voisinages forts, qui constitue notre sélection pour le groupe à 5 et 10 positions actives en utilisant un seuil égal à 10, ils sont dans les tableaux 4.12 et 4.13. Pour les sélections 20-actives et 30-actives, il a fallu descendre le seuil à 1 pour obtenir les 45 voisinages forts proches de l'objectif, il manque une ou deux positions dans quelques cas, voir les tables 4.14 et 4.15.

Table 4.12 – La sélection des positions pour tests 5-actives

Nom	positions actives
1A81 1	10 13 16 84 86
1A81 2	20 21 24 27 116
1A81 3	35 38 56 105 107
1A81 4	44 47 52 65 67
1A81 5	82 84 86 87 90
1ABO 1	64 66 90 93 100
1ABO 2	72 74 80 104 111
1ABO 3	79 82 102 111 115
1ABO 4	83 86 104 105 106
1ABO 5	93 100 102 113 116
1BM2 1	101 106 140 141 146
1BM2 2	120 128 131 132 135
1BM2 3	58 61 127 128 129
1BM2 4	74 75 98 100 105
1BM2 5	85 87 95 110 128
1CKA 1	136 138 158 175 190
1CKA 2	149 166 169 171 181
1CKA 3	151 153 157 159 172
1CKA 4	164 170 172 184 187
1CKA 5	172 174 182 186 187
1G9O 1	10 13 54 57 92
1G9O 2	15 39 42 54 57
1G9O 3	24 26 28 39 42
1G9O 4	48 53 57 59 88
1G9O 5	75 78 79 86 88
1M61 1	12 20 23 24 27
1M61 2	17 20 24 37 49
1M61 3	27 33 51 100 102
1M61 4	5 8 10 11 36
1M61 5	59 71 84 87 94
1O4C 1	20 21 32 34 46
1O4C 2	2 71 79 81 82
1O4C 3	33 45 63 71 73
1O4C 4	43 45 63 71 85
1O4C 5	8 33 82 83 86
1R6J 1	194 237 239 270 272
1R6J 2	199 201 211 218 232
1R6J 3	213 218 227 232 238
1R6J 4	221 227 232 267 269
1R6J 5	241 254 258 267 269
2BYG 1	189 191 221 244 246
2BYG 2	205 224 239 245 248
2BYG 3	232 233 265 272 274
2BYG 4	238 240 243 276 278
2BYG 5	253 261 264 265 274

Table 4.13 – La sélection des positions pour tests 10-actives

Nom	positions actives
1A81 1	13 15 39 41 53 86 89 90 93 103
1A81 2	39 41 53 55 64 66 76 89 92 103
1A81 3	51 53 64 66 68 74 76 82 88 92
1A81 4	76 82 87 88 90 91 92 95 97 99
1A81 5	9 10 11 16 41 51 53 66 88 89
1ABO 1	64 72 74 79 89 91 101 103 108 111
1ABO 2	66 68 80 82 88 90 100 102 104 111
1ABO 3	69 70 72 74 80 81 106 113 114 115
1ABO 4	71 78 83 84 94 99 101 104 105 106
1ABO 5	72 79 82 94 99 102 104 106 111 115
1BM2 1	119 120 121 122 123 125 131 134 135 140
1BM2 2	125 126 127 129 130 133 134 136 137 147
1BM2 3	83 99 101 106 108 135 140 141 146 148
1BM2 4	85 95 97 110 118 120 125 128 131 132
1BM2 5	99 101 106 139 140 141 142 143 144 146
1CKA 1	134 135 160 161 162 173 174 175 176 179
1CKA 2	137 139 143 151 153 157 159 172 182 186
1CKA 3	138 140 147 149 150 155 166 169 181 188
1CKA 4	140 141 153 154 155 157 174 175 184 186
1CKA 5	151 153 157 166 168 173 174 176 178 179
1G9O 1	10 11 13 14 15 16 53 54 57 92
1G9O 2	15 17 24 26 39 42 48 51 53 88
1G9O 3	26 28 39 42 48 53 57 59 88 90
1G9O 4	34 35 58 60 68 70 74 75 89 91
1G9O 5	71 73 74 77 80 81 82 83 84 85
1M61 1	10 12 20 23 24 27 35 49 102 104
1M61 2	17 20 21 24 37 39 40 47 49 58
1M61 3	34 36 46 48 59 61 71 83 84 87
1M61 4	5 6 11 36 46 48 61 69 83 84
1M61 5	59 61 70 71 75 77 83 86 87 92
1O4C 1	31 33 45 47 61 63 73 86 89 100
1O4C 2	50 51 52 53 63 72 73 77 85 89
1O4C 3	61 62 63 71 72 73 79 85 88 89
1O4C 4	73 74 75 76 77 89 92 94 96 101
1O4C 5	90 91 93 96 98 99 101 102 103 104
1R6J 1	193 194 195 197 199 218 232 236 267 269
1R6J 2	199 209 211 213 218 227 232 238 265 267
1R6J 3	201 204 205 209 211 218 241 258 265 267
1R6J 4	209 211 213 218 227 238 241 258 265 267
1R6J 5	238 240 241 242 246 257 258 261 265 267
2BYG 1	194 196 203 205 224 233 239 245 274 276
2BYG 2	203 205 207 224 227 233 239 243 245 276
2BYG 3	206 207 222 245 248 253 256 261 264 265
2BYG 4	221 222 245 248 251 253 256 261 264 265
2BYG 5	247 248 249 250 251 252 259 262 263 275

Table 4.14 – La sélection des positions pour tests 20-actives

Nom	positions actives
1A81 1	9 11 12 13 15 16 17 19 20 25 28 39 40 41 42 43 44 45 114 117
1A81 2	9 11 12 13 15 16 17 19 20 25 28 39 40 41 42 43 44 45 47 117
1A81 3	9 11 12 13 15 16 17 19 41 43 48 51 68 74 84 86 109 114 117
1A81 4	12 13 15 16 17 19 20 25 28 39 40 41 42 43 44 45 47 86 114 117
1A81 5	13 15 16 19 41 43 48 51 60 64 68 70 71 74 84 86 87 88 109 114 117
1ABO 1	64 66 67 68 82 86 87 88 89 90 91 101 102 103 108 111 113 116
1ABO 2	64 65 66 67 84 87 88 89 90 91 93 100 101 102 103 108 111 113 116
1ABO 3	65 66 67 87 88 89 90 91 93 94 95 100 101 102 103 106 108 111 113 116
1ABO 4	64 65 66 67 69 87 88 89 90 91 93 100 101 102 103 106 108 111 113 116
1ABO 5	66 67 68 82 86 87 88 89 90 91 93 100 101 102 103 106 108 111 113 116
1BM2 1	55 56 60 61 62 83 84 85 86 87 95 97 99 110 118 125 127 133 150 152
1BM2 2	55 56 60 61 62 83 84 85 86 87 95 97 99 110 118 125 127 128 129 152
1BM2 3	55 56 58 60 61 62 64 67 69 73 83 84 85 86 87 129 132 133 150 152
1BM2 4	55 56 60 61 62 69 83 84 85 86 87 95 97 99 110 129 132 133 150 152
1BM2 5	58 60 60 61 61 62 64 67 69 73 75 83 84 85 86 129 132 133 150 152
1CKA 1	134 135 136 137 138 139 150 151 160 161 162 163 164 170 171 172 173 179 189 190
1CKA 2	134 135 136 137 139 150 151 153 160 161 162 163 164 170 171 172 173 179 189 190
1CKA 3	134 136 137 139 150 151 157 158 160 161 162 163 164 170 171 172 173 179 189 190
1CKA 4	136 137 139 150 151 153 158 159 160 161 162 163 164 170 171 172 173 179 189 190
1CKA 5	137 139 150 151 153 158 160 161 162 163 164 170 171 172 173 174 175 179 189 190
1G9O 1	9 10 11 13 14 15 31 34 38 54 57 58 60 68 90 91 92 94 95 96
1G9O 2	9 11 13 14 15 16 31 34 38 54 57 58 60 68 90 91 92 94 95 96
1G9O 3	9 11 13 14 15 31 34 38 54 55 57 58 60 68 90 91 92 94 95 96
1G9O 4	9 11 13 15 16 17 54 57 58 59 60 61 68 89 90 91 92 94 95 96
1G9O 5	10 11 13 15 16 17 54 57 58 60 61 68 89 90 91 92 94 95 96
1M61 1	34 35 36 37 38 39 42 46 47 48 49 50 61 63 69 71 77 78 81 82
1M61 2	35 36 37 38 39 42 46 47 48 49 50 61 63 69 71 77 78 81 82 83
1M61 3	38 39 42 46 47 48 49 50 61 63 69 71 77 78 81 82 83 84 85 87
1M61 4	42 46 47 48 49 50 61 63 69 71 77 78 81 82 83 84 85 87 88 98
1M61 5	5 7 50 61 63 69 71 77 78 81 82 83 84 85 87 88 98 103 104 109
1O4C 1	32 33 34 35 43 45 63 65 71 73 79 80 81 82 83 84 85 86 87 89
1O4C 2	3 4 5 7 8 9 11 17 31 32 33 34 35 43 45 63 65 71 73 79
1O4C 3	1 3 4 5 6 7 8 9 11 17 31 32 33 35 43 45 65 81 82 83
1O4C 4	1 2 3 4 5 6 7 8 9 11 12 13 14 17 19 35 65 81 82 83
1O4C 5	1 3 4 5 6 7 8 9 11 12 17 31 32 33 34 35 65 81 82 83
1R6J 1	193 194 195 197 214 215 217 218 233 235 236 237 239 240 241 242 247 269 270 273
1R6J 2	193 194 197 198 199 217 233 235 236 237 238 239 240 241 242 247 268 270 272 273
1R6J 3	193 195 197 217 233 235 236 239 240 241 242 244 245 247 268 269 270 272 273
1R6J 4	193 195 197 217 233 235 236 237 239 241 242 244 245 247 268 269 270 272 273
1R6J 5	193 194 197 198 199 233 236 237 239 240 241 247 268 269 270 272 273
2BYG 1	186 187 188 189 190 191 192 215 216 219 244 246 270 271 273 274 278 280 281 282
2BYG 2	186 187 188 189 190 215 216 219 221 223 240 243 270 271 273 274 278 280 281 282
2BYG 3	186 187 188 189 190 215 216 219 221 223 240 243 244 270 271 273 278 280 281 282
2BYG 4	186 187 188 189 190 215 216 219 221 223 240 243 244 270 274 276 278 280 281 282
2BYG 5	187 189 190 191 192 215 216 219 221 223 240 243 244 270 274 276 278 280 281 282

Table 4.15 – La sélection des positions pour tests 30-actives

Nom	positions actives
1A81 1	9 11 12 13 15 16 17 19 20 25 26 27 28 29 36 38 39 40 41 42 43 48 51 68 74 84 86 109 114 117
1A81 2	9 10 11 12 13 15 16 17 19 20 25 28 39 41 43 48 51 68 74 83 84 86 87 88 90 91 93 109 114 117
1A81 3	9 11 12 13 15 16 17 19 20 25 27 28 36 38 39 40 41 42 43 48 51 68 74 84 86 109 114 117
1A81 4	9 10 11 12 13 15 16 17 19 20 25 28 36 39 40 41 42 43 44 45 48 51 68 74 84 86 109 114 117
1A81 5	9 10 11 12 13 15 16 17 19 20 25 28 39 40 41 42 43 44 45 47 48 51 52 68 74 84 86 109 114 117
1ABO 1	64 65 66 67 68 70 71 72 75 78 79 80 81 82 83 86 87 88 89 90 91 93 100 101 102 103 108 111 113 116
1ABO 2	64 65 66 67 68 72 75 78 80 81 82 83 84 86 87 88 89 90 91 93 94 100 101 102 103 104 108 111 113 116
1ABO 3	64 66 67 68 70 71 72 78 82 86 87 88 89 90 91 93 94 95 96 99 100 101 102 103 104 105 108 111 113 116
1ABO 4	64 65 66 67 70 71 72 68 82 86 87 88 89 90 91 93 94 95 96 99 100 101 102 103 104 105 108 111 113 116
1ABO 5	65 66 67 70 71 72 75 78 80 82 86 87 88 89 90 91 93 94 95 96 99 100 101 102 103 108 111 113 116
1BM2 1	55 56 58 60 61 62 83 84 85 86 87 95 97 99 110 118 119 120 121 122 125 127 128 129 130 131 132 133 150 152
1BM2 2	56 58 60 61 62 83 84 85 86 87 95 97 99 110 118 119 120 121 122 123 125 127 128 129 130 131 132 133 150 152
1BM2 3	58 60 61 62 83 84 85 86 87 95 97 99 108 109 110 118 120 121 122 123 125 127 128 129 132 133 134 135 150 152
1BM2 4	55 56 58 60 61 62 83 84 85 86 87 95 97 99 108 109 110 118 120 121 125 127 128 129 130 131 132 133 150 152
1BM2 5	56 58 60 61 62 67 83 84 85 86 87 95 97 99 110 111 112 113 115 118 125 127 128 129 130 131 132 133 150 152
1CKA 1	134 135 136 137 139 140 141 142 143 144 146 147 148 149 150 151 158 159 160 161 162 163 164 170 171 172 173 179 189 190
1CKA 2	134 135 136 137 139 143 144 146 147 148 149 150 151 158 159 160 161 162 163 164 170 171 172 173 179 186 187 188 189 190
1CKA 3	135 136 137 139 144 146 147 148 149 150 151 157 158 159 160 161 162 163 164 170 171 172 173 179 186 187 188 189 190
1CKA 4	136 137 139 140 141 142 143 144 146 147 148 151 158 159 160 161 162 163 164 170 171 172 173 179 184 186 187 188 189 190
1CKA 5	134 136 137 139 140 141 142 143 144 146 147 148 151 158 159 160 161 162 163 164 170 171 172 173 179 182 187 188 189 190
1G9O 1	9 10 11 13 15 24 31 34 38 40 41 42 43 46 48 49 50 51 54 57 58 60 68 89 90 91 92 94 95 96
1G9O 2	9 11 13 15 31 32 34 38 40 41 42 43 46 48 49 50 51 54 57 58 60 68 89 90 91 92 94 95 96
1G9O 3	9 10 11 13 15 31 32 34 38 40 41 42 43 46 48 49 50 51 54 57 58 60 68 89 90 91 92 94 95 96
1G9O 4	10 11 13 14 15 31 32 34 38 40 41 42 43 46 48 49 50 51 54 57 58 60 61 68 89 90 91 92 94 95 96
1G9O 5	10 11 13 14 15 31 32 40 41 42 43 46 48 49 50 51 54 57 58 60 61 62 68 87 89 90 91 92 94 95 96
1M61 1	12 14 15 20 34 35 36 37 38 39 42 46 47 48 49 50 61 63 69 71 77 78 81 82 83 84 85 87 88 98
1M61 2	6 7 8 10 11 12 14 15 20 21 34 35 36 37 38 39 42 46 47 48 49 50 61 63 69 71 77 78 81 82
1M61 3	5 7 35 36 37 38 39 42 46 47 48 49 50 61 63 69 71 77 78 81 82 83 84 85 87 88 98 103 104 109
1M61 4	7 8 10 11 12 14 15 20 34 35 36 37 38 39 42 46 47 48 49 50 61 63 69 71 77 78 81 82 83 84
1M61 5	8 10 11 12 14 15 20 34 35 36 37 38 39 42 46 47 48 49 50 61 63 69 71 77 78 81 82 83 84 85
1O4C 1	1 2 3 4 5 6 7 8 9 11 17 31 32 33 34 35 43 45 63 65 71 81 82 83 90 91 92 93 94 96
1O4C 2	1 3 4 5 7 8 9 11 17 31 32 33 34 35 43 45 63 65 71 73 79 80 81 82 83 90 91 92 93 96
1O4C 3	1 3 4 5 6 7 8 9 11 17 31 32 33 34 35 43 45 63 65 71 73 79 80 81 82 83 90 91 92 93 96
1O4C 4	1 3 4 5 7 8 9 11 17 31 32 33 34 35 43 45 63 65 71 73 79 80 81 82 83 84 91 92 93 96
1O4C 5	1 3 4 5 6 7 8 9 11 17 31 32 33 34 35 43 45 63 65 71 73 79 80 81 82 83 84 92 93 96
1R6J 1	193 194 195 197 198 199 217 218 219 220 221 222 223 224 225 226 227 228 229 233 235 236 237 239 247 268 269 270 272 273
1R6J 2	193 194 195 197 198 199 217 220 221 222 223 224 225 226 227 228 229 233 235 236 237 239 240 247 268 269 270 272 273
1R6J 3	193 194 195 197 198 199 208 217 220 221 222 223 224 225 226 227 228 229 230 233 235 236 237 239 247 268 269 270 272 273
1R6J 4	193 194 195 197 198 199 217 220 221 222 223 224 225 226 227 228 229 233 235 236 237 239 240 247 249 268 269 270 272 273
1R6J 5	194 195 197 198 199 217 220 221 222 223 224 225 226 227 228 229 233 235 236 237 239 240 247 249 250 268 269 270 272 273
2BYG 1	186 187 188 189 190 191 192 215 216 219 221 223 240 243 244 246 250 251 252 253 254 255 256 257 259 260 278 280 281 282
2BYG 2	186 187 188 189 190 191 192 198 215 216 219 221 223 240 243 244 251 252 253 254 255 256 257 259 260 278 280 281 282
2BYG 3	186 187 188 189 190 191 192 215 216 219 221 223 240 243 244 251 252 253 254 255 256 257 259 260 278 246 280 281 282
2BYG 4	186 187 188 189 190 191 192 215 216 219 221 223 240 243 244 245 246 251 252 253 254 255 256 257 259 260 278 281 282
2BYG 5	186 187 188 189 190 191 192 215 216 219 221 223 240 243 244 246 251 252 253 254 255 256 257 259 260 278 280 281 282

Chapitre 5

PDZ

5.1 Introduction

Nous cherchons maintenant, à évaluer la performance de notre modèle CPD sur un ensemble de protéines. Les domaines PDZ (« Postsynaptic density-95 / Discs large / Zonula occludens-1 ») sont de petits domaines globulaires qui établissent des réseaux d’interactions entre protéines dans la cellule [98–102]. Ils forment des interactions spécifiques avec des protéines cibles, généralement en reconnaissant quelques acides aminés à l’extrémité C-terminale. En raison de leur importance biologique, les domaines PDZ et leur interaction avec les protéines cibles ont été largement étudiés et utilisés en conception *in silico*. Des ligands ont été conçus pour moduler l’activité de domaines PDZ impliqués dans diverses pathologies [103, 104]. Des domaines PDZ et leurs ligands redessinés ont été utilisés pour élucider les principes du repliement des protéines et de l’évolution [105–107]. Ainsi, ces domaines avec leurs ligands peptidiques fournissent des « benchmarks » pour tester les méthodes informatiques elles-mêmes [108, 109].

À partir d’une sélection de domaines PDZ, nous optimisons les énergies de référence, paramètres essentiels dans notre modèle, grâce à la méthode du maximum de vraisemblance. La performance du modèle est testée en générant des séquences par Proteus pour chaque protéine de la sélection. Pour cela, nous nous basons sur les résultats précédents, en particulier ceux de la section 4.3.2, pour définir les valeurs des paramètres du Monte Carlo. Nous confrontons nos résultats à ceux de la fonction d’énergie de Rosetta, fonction qui a connu le plus de succès et qui est la plus souvent citée [110]. Elle comprend un terme de répulsion de Lennard-Jones, un terme Coulomb, un terme de liaison hydrogène, un terme de solvatation Lazaridis-Karplus et des énergies de référence d’état dépliées, mais est plus empirique que la nôtre. Il y a un grand nombre de paramètres spécifiquement optimisés pour le CPD, qui offrent des performances optimales, mais une interprétation physique moins transparente que Proteus.

La production de nos séquences calculées est effectuée par des simulations Monte Carlo où toutes les positions de la chaîne polypeptidique sont autorisées à muter librement, excepté celles occupées par une glycine ou une proline qui conservent leur type d'acide aminé. Ces exceptions découlent de la contrainte du backbone fixe, puisque ces deux types d'acides aminés influent sur la structure du squelette. Nous obtenons ainsi des milliers de variantes pour chaque domaine étudié. Nos tests comprennent une validation directe et une validation croisée dans laquelle les énergies de référence optimisées sur un sous-ensemble de notre sélection sont utilisées sur un autre sous-ensemble de cette même sélection.

Nous réalisons également une série de simulations Monte Carlo de deux domaines PDZ où le potentiel chimique hydrophobe des types d'acides aminés est progressivement augmenté, polarisant artificiellement la composition de la protéine. Comme le biais hydrophobe augmente, les acides aminés hydrophobes envahissent progressivement la protéine de l'intérieur. La propension de chaque position du noyau à devenir hydrophobe à un niveau de biais plus ou moins élevé peut être considérée comme un indice d'hydrophobicité. Il dépend de la structure et nous renseigne sur la capacité de la protéine à supporter des mutations.

5.2 Le modèle d'état déplié

5.2.1 Les énergies de référence

Nous travaillons avec l'énergie de repliement de la protéine, c'est-à-dire la différence entre son énergie à l'état replié et son énergie à l'état déplié. Un mouvement élémentaire possible est une « mutation ». Il s'agit d'une modification du type de chaîne latérale à une position choisie i dans la protéine pliée, en assignant un rotamère particulier à la nouvelle chaîne latérale et de la modification inverse dans l'état déplié, voir les détails en 2.1. Pour une séquence S particulière, nous avons introduit à la section 1.1.1 l'énergie de l'état déplié comme :

$$E^u = \sum_{i \in S} E_{t_i}^r \quad (5.1)$$

ici, i est la position dans la séquence et t_i le type en i , la somme se faisant sur tous les résidus de S . Les grandeurs E_t^r sont les « énergies de référence ». Ce sont des paramètres essentiels dans notre modèle de simulation.

5.2.2 La vraisemblance des énergies de référence

Notre objectif maintenant est de déterminer les énergies de référence empiriquement afin que les simulations produisent des fréquences d'acides aminés qui correspondent à un ensemble de valeurs cibles, notamment des valeurs expérimentales. Pour cela, nous choisissons les E_t^r qui maximisent la probabilité de Boltzmann d'un ensemble de séquences expérimentales. C'est à dire, nous retenons les E_t^r les plus vraisemblables étant donnée l'observation des séquences expérimentales. Soit S une séquence particulière. Sa probabilité de Boltzmann est :

$$p(S) = \frac{1}{Z} \exp(-\beta \Delta G_S) \quad (5.2)$$

avec

$$\Delta G_S = G_S^f - E_S^u \quad (5.3)$$

ΔG_S est l'énergie libre de repliement de S , G_S^f est l'énergie libre de l'état replié, $\beta = \frac{1}{kT}$ est la température inverse et Z une constante de normalisation (la fonction de partition). En injectant 5.3 dans 5.2, nous avons alors :

$$kT \ln p(S) = \sum_{i \in S} E^r(t_i) - G_S^f - kT \ln Z = \sum_{t \in aa} n_S(t) E_t^r - G_S^f - kT \ln Z \quad (5.4)$$

la somme à droite se fait sur l'ensemble des types d'acides aminés et $n_S(t)$ est le nombre d'acides aminés de type t dans S .

Nous considérons maintenant un ensemble \mathcal{S} de N séquences cibles ; on appelle \mathcal{L} la probabilité d'observer l'ensemble \mathcal{S} en entier. \mathcal{L} est fonction des paramètres du modèle E_t^r . Comme nous voulons le maximum de \mathcal{L} sur les E_t^r , nous nous référerons à \mathcal{L} comme la vraisemblance des E_t^r [111]. Nous avons :

$$kT \ln \mathcal{L} = \sum_S \sum_{i \in aa} n_S(t) E_t^r - \sum_S G_S^f - N kT \ln Z = \sum_{t \in aa} N(t) E_t^r - \sum_S G_S^f - N kT \ln Z \quad (5.5)$$

avec $N(t)$ le nombre d'acides aminés de type t dans l'ensemble \mathcal{S} . Le facteur de normalisation Z est une somme sur l'ensemble les séquences possibles \mathcal{R} :

$$Z = \sum_{\mathcal{R}} \exp(-\beta \Delta G_{\mathcal{R}}) = \sum_{\mathcal{R}} \exp(-\beta \Delta G_{\mathcal{R}}^f) \Pi_{t \in aa} \exp(\beta n_{\mathcal{R}}(t) E_t^r) \quad (5.6)$$

Pour maximiser \mathcal{L} , nous considérons la dérivé de Z selon chacune des E_t :

$$\frac{\partial Z}{\partial E_t^r} = \sum_{\mathcal{R}} \beta n_{\mathcal{R}}(t) \exp(-\beta \Delta G_{\mathcal{R}}^f) \Pi_{s \in aa} \exp(\beta n_{\mathcal{R}}(s) E_s^r) \quad (5.7)$$

Nous avons alors :

$$\frac{kT}{Z} \frac{\partial Z}{\partial E_t^r} = \frac{\sum_{\mathcal{R}} n_{\mathcal{R}}(t) \exp(-\beta \Delta G_{\mathcal{R}})}{\sum_{\mathcal{R}} \exp(-\beta \Delta G_{\mathcal{R}})} = \langle n(t) \rangle. \quad (5.8)$$

La quantité à droite est la moyenne de Boltzmann du nombre $n(t)$ des acides aminés de type t sur toutes les séquences possibles. Comme une simulation Monte Carlo converge vers la distribution de Boltzmann, la population moyenne de t que nous obtenons dans nos simulations converge vers la moyenne de Boltzmann de $n(t)$. En pratique, nous obtenons cette quantité comme la population moyenne de t dans une simulation Monte Carlo assez longue.

Pour que $\ln \mathcal{L}$ soit maximal il faut que ses dérivées par rapport à E_t^r soient nulles.

$$\frac{1}{N} \frac{\partial}{\partial E_t^r} \ln \mathcal{L} = \frac{1}{N} \sum_S n_S(t) - \langle n(t) \rangle = \frac{N(t)}{N} - \langle n(t) \rangle \quad (5.9)$$

et donc

$$\mathcal{L} \text{ maximum} \implies \frac{N(t)}{N} = \langle n(t) \rangle, \forall t \in \text{aa}$$

Ainsi, pour maximiser \mathcal{L} , nous choisissons les E_t^r tels qu'une longue simulation donne les mêmes fréquences d'acides aminés que l'ensemble cible.

5.2.3 Recherche du maximum de vraisemblance

Nous utilisons trois méthodes pour approcher les valeurs E_t^r .

- La première consiste à avancer dans la direction du gradient de $\ln(\mathcal{L})$ en utilisant la règle itérative suivante [111], nous l'appelons méthode linéaire :

$$E_t^r(n+1) = E_t^r(n) + \alpha \frac{\partial}{\partial E_t^r} \ln(\mathcal{L}) = E_t^r(n) + \delta E (n_t^{exp} - \langle n(t) \rangle_n) \quad (5.10)$$

avec α une constante, $n_t^{exp} = \frac{N(t)}{N}$ la population moyenne d'acide aminé de type t dans l'ensemble ciblé, $\langle \cdot \rangle_n$ indique une moyenne sur une simulation effectuée en utilisant les énergies de références courantes $E_t^r(n)$, et δE une constante empirique avec la dimension d'une énergie, correspondant à l'amplitude de mise à jour. Cette procédure de mise à jour est répétée jusqu'à convergence.

- La deuxième méthode est une variante de la première dans laquelle le δE n'est pas constant, mais ajusté au cours de la simulation de la façon suivante. On introduit une fonction proxy C comme outil de mesure rapide de l'état de l'optimisation, de

la façon suivante :

$$C = \sum_{t \in aa} (n_t^{exp} - \langle n(t) \rangle_n)^2 \quad (5.11)$$

Alors, la règle 5.10 est utilisée trois fois avec trois valeurs différentes pour le δE ceci avec un jeu d'énergie de références identiques. Une interpolation parabolique est effectuée sur les trois valeurs de la fonction C obtenues, le minimum de la parabole est calculé et est utilisé comme δE pour le cycle suivant, au terme duquel les énergies sont mises à jour.

- La troisième méthode, utilisée précédemment [91, 92], utilise une règle de mise à jour logarithmique :

$$E_t^r(n+1) = E_t^r(n) + kT \ln \frac{\langle n(t) \rangle_n}{n_t^{exp}} \quad (5.12)$$

avec kT l'énergie thermique, fixée empiriquement à 0,5 kcal/mol. Nous l'appelons la méthode logarithmique. Dans les dernières itérations, certaines valeurs ont tendance à converger lentement, avec des oscillations. Par conséquent, une règle modifiée est ajoutée dans laquelle une énergie au cycle n et l'énergie au cycle $n-1$ sont moyennées avec un poids respectif de 2/3 et 1/3 afin d'atténuer les vibrations.

5.3 Méthodes de calcul

5.3.1 Fonction énergétique efficace pour l'état replié

La matrice énergétique d'une protéine est calculée avec la fonction d'énergie suivante :

$$E = E_{MM} + E_{GB} + \sum_i \sigma_i A_i \quad (5.13)$$

E_{MM} représente l'énergie interne de la protéine et se compose de six termes détaillés en 1.2.1. Les autres termes de l'équation 5.13 représentent la contribution du solvant. Nous utilisons un modèle de solvant implicite « Generalized Born + Surface Area » ou GBSA (1.3.5 et 1.12). Ici A_i est la surface exposée accessible au solvant de l'atome i , σ_i est un paramètre qui représente la préférence de chaque atome à être exposé ou caché du solvant. Les atomes du soluté sont divisés en quatre groupes avec pour chacun une valeur σ_i spécifique en cal/mol/Å². Nous travaillons avec deux jeux différents pour les σ_i . Le premier et un jeu déjà utilisé dans différentes études [112, 113] : non polaire -5, aromatique -40, polaire -80, ionique -100. Le second provient d'une étude plus récente faite dans le

notre laboratoire, dans laquelle il a été optimisé sur un ensemble de protéines PDZ [59] : non polaire -3, aromatique -1, polaire -9, ionique -9. Dans les deux cas, on attribue aux atomes d'hydrogène un coefficient de surface de 0.

Les surfaces sont calculées par l'algorithme de Lee et Richards [96], qui est implémenté dans le programme Xplor [49] et expliqué en 2.3.1. Les simulations MC utilisent une constante diélectrique pour la protéine $\epsilon_p = 4$ ou 8 (voir la partie Résultats). Nous utilisons deux variantes du modèle GB, la méthode NEA pour « Native Environment Approximation » et la méthode FDB « Fluctuating Dielectric Boundary » [51]. Elles sont présentées respectivement en 1.3.5 et 2.3.3.

Le champ de force utilisé, Amber ff99SB [19], est légèrement modifié pour le CPD, en remplaçant les charges du backbone par un ensemble unifié, obtenu en faisant la moyenne sur l'ensemble des types d'acides aminés et ajustant légèrement pour rendre la partie backbone de chaque acide aminé neutre [114].

5.3.2 Les énergies de référence de l'état déplié

Dans le modèle CPD, l'énergie de l'état déplié dépend de la composition de la séquence par l'ensemble des énergies de référence E_t^r (équation 5.1). Ici, les énergies de référence sont attribuées en fonction des types d'acides aminés t , mais aussi de la position de chaque acide aminé dans la structure repliée à travers son caractère enfoui ou exposé au solvant. Ainsi, pour un type donné, il y a deux valeurs distinctes de E_t^r , une enfouie et une exposée. Cette approche se justifie par les trois éléments suivants :

- Nous supposons que la structure résiduelle est présente dans l'état déplié, de sorte que les acides aminés conservent en partie leur caractère enfoui/exposé.
- Nous supposons que le modèle d'état déplié compense de manière systématique des erreurs dans la fonction d'énergie de l'état plié, de sorte que la structure pliée contribue indirectement aux énergies de référence.
- Cette stratégie rend le modèle moins sensible aux variations de la longueur des boucles de surface et au ratio du nombre de résidus de surface sur celui des enfouis, qui peut varier considérablement selon les homologues (voir plus bas).

Par conséquent, ce modèle devrait être plus transférable à l'intérieur d'une famille de protéines. Distinguer les positions enfouies/exposées double le nombre de paramètres E_t^r à ajuster. Inversement, pour réduire le nombre de paramètres, nous groupons les acides

aminés en classes homologues, voir table 5.1. Dans chaque classe c, et pour chaque type de position (enfoui ou exposé), les énergies de référence ont la forme :

$$E_t^r = E_c^r + \delta E_t^r \quad (5.14)$$

E_c^r est un paramètre ajustable, tandis que δE_t^r est une constante, calculée comme la différence d'énergie de mécanique moléculaire entre les types d'acides aminés de classe c, supposée en conformation dépliée où chaque acide aminé interagit uniquement avec lui-même et avec le solvant.

Groupe	acides aminés	propriétés
1	Ala, Cys, Thr	petit
2	Ser	
3	Glu, Asp	chargé négativement
4	Gln, Asn	polaire
5	Ile, Leu, Val	apolaire
6	Met	non polaire
7	Hip, Hid, Hie	chargé positivement
8	Arg	
9	Lys	
10	Phe, Trp	aromatique
11	Tyr	
12	Gly, Pro	non mutable

Table 5.1 – Les groupes d'acides aminés utilisés pour l'optimisation des énergies de référence.

Plus précisément, nous effectuons des simulations MC d'un peptide étendu (le peptide *Syndecan1*) et calculons les énergies moyennes pour chaque type d'acide aminé à chaque position peptidique (à l'exclusion des positions terminales). Nous prenons les différences entre les types d'acides aminés et les moyennons sur les positions peptidiques. Pendant, la maximisation de la vraisemblance, E_c^r est optimisé tandis que δE_t^r est fixe. Pour optimiser les valeurs E_t^r , nous utilisons une des trois méthodes de la section 5.2.3 avec comme fréquences cibles les fréquences expérimentales soit des classes d'acides aminés, soit des types d'acides aminés. Le début d'optimisation se fait sur les classes, puis lorsque la convergence est correctement établie, c'est-à-dire lorsque la fonction proxy C calculée sur ces classes retourne des valeurs stables et faibles, nous relâchons cette contrainte pour

optimiser sur l'ensemble des types d'acide aminé mutables. Typiquement, vingt cycles d'optimisation sont effectués sur les classes, puis encore vingt cycles sur les types.

5.4 Séquences expérimentales et modèles structuraux

5.4.1 L'ensemble des protéines PDZ

Nous sélectionnons huit protéines de la famille PDZ dont les structures cristallographiques sont connues. Aux trois présentes dans l'ensemble étudié au chapitre précédent, NHREF, Syntenin et DLG2, sont ajoutées les protéines INAD, GRIP, PSD95, Cask et Tiam1. Leur séquence est présentée à la figure 5.1. Le nombre de positions actives, c'est-à-dire les positions qui vont être mutées, est du même ordre pour chaque protéine (voir le tableau 5.2).

Table 5.2 – La sélection de domaines protéiques PDZ

nom	Code PDB	résidus	nombre de positions actives
NHREF	1G9O	9-99	76
INAD	1IHJ	13-105	82
GRIP	1N7E	668-761	79
Syntenin	1R6J	193-273	72
DLG2	2BYG	186-282	82
PSD95	3K82	305-402	80
Cask	1KWA	487-568	74
Tiam1	4GVD	838-930	84

5.4.2 Alignements Blast croisés

Pour caractériser les homologies dans cet ensemble, une série de requêtes BLAST est effectuée sur chaque paire de séquences en utilisant le programme `blastp` avec les options comme indiqué en 2.7.6. Il apparaît que Syntenin et Tiam1 sont atypiques dans l'ensemble : il n'y a pas d'homologues avec une E-value inférieure à 10^{-7} et plusieurs E-value supérieur à 10. PSD95 est la protéine la plus consensuelle, ayant d'une part une homologie avec toutes les autres à au plus $6 \cdot 10^{-4}$, et d'autre part ayant quatre homologues à moins de $2 \cdot 10^{-10}$, pour un pourcentage d'identité compris entre 30 et 46. Globalement, il n'y a que peu d'homologies, la plus forte n'étant que de $3 \cdot 10^{-15}$ entre PSD95 et DLG2 pour un pourcentage d'identité de 37. Les détails sont dans le tableau 5.3.

Table 5.3 – E-value et pourcentage d’identité des alignements Blast native versus native pour nos séquences PDZ

Proteine	NHREF	INAD	GRIP	Syntenin	DLG2	PSD95	CASK	TIAM1
NHREF	$2 \cdot 10^{-66}$ (100)	$5 \cdot 10^{-10}$ (40)	$2 \cdot 10^{-3}$ (25)	$3 \cdot 10^{-7}$ (25)	$2 \cdot 10^{-11}$ (35)	$1 \cdot 10^{-12}$ (30)	$5 \cdot 10^{-5}$ (25)	$9 \cdot 10^{-7}$ (35)
INAD	$5 \cdot 10^{-10}$ (40)	$3 \cdot 10^{-68}$ (100)	$2 \cdot 10^{-7}$ (27)	[18]	$2 \cdot 10^{-8}$ (27)	$9 \cdot 10^{-14}$ (46)	$4 \cdot 10^{-6}$ (35)	[16]
GRIP	$2 \cdot 10^{-3}$ (25)	$2 \cdot 10^{-7}$ (27)	$3 \cdot 10^{-67}$ (100)	[21]	$3 \cdot 10^{-14}$ (36)	$2 \cdot 10^{-10}$ (37)	$9 \cdot 10^{-12}$ (30)	$5 \cdot 10^{-5}$ (35)
Syntenin	$3 \cdot 10^{-7}$ (25)	[18]	[21]	$1 \cdot 10^{-59}$ (100)	[17]	$1 \cdot 10^{-6}$ (32)	$7 \cdot 10^{-3}$ (32)	[18]
DLG2	$2 \cdot 10^{-11}$ (35)	$2 \cdot 10^{-8}$ (27)	$3 \cdot 10^{-14}$ (37)	[17]	$7 \cdot 10^{-71}$ (100)	$3 \cdot 10^{-15}$ (37)	$2 \cdot 10^{-7}$ (28)	$5 \cdot 10^{-5}$ (41)
PSD95	$1 \cdot 10^{-12}$ (30)	$9 \cdot 10^{-14}$ (46)	$2 \cdot 10^{-10}$ (36)	$1 \cdot 10^{-6}$ (32)	$3 \cdot 10^{-15}$ (37)	$4 \cdot 10^{-70}$ (100)	$1 \cdot 10^{-7}$ (27)	$6 \cdot 10^{-4}$ (33)
Cask	$5 \cdot 10^{-5}$ (25)	$4 \cdot 10^{-6}$ (35)	$9 \cdot 10^{-12}$ (30)	$7 \cdot 10^{-3}$ (32)	$2 \cdot 10^{-7}$ (28)	$1 \cdot 10^{-7}$ (27)	$7 \cdot 10^{-61}$ (100)	$5 \cdot 10^{-4}$ (33)
Tiam1	$9 \cdot 10^{-7}$ (35)	[16]	$5 \cdot 10^{-5}$ (35)	[18]	$5 \cdot 10^{-5}$ (41)	$6 \cdot 10^{-4}$ (33)	$5 \cdot 10^{-4}$ (33)	$1 \cdot 10^{-68}$ (100)

S'il n'y a pas de touche avec une E-value inférieure à 10, [.] donne le pourcentage d'identité du couple dans l'alignement des six séquences sauvages.

5.4.3 Sélection des homologues

Pour définir les fréquences d'acides aminés cibles nécessaires pour maximiser nos vraisemblances, nous sélectionnons un ensemble de séquences homologues pour chacune des six premières protéines de notre sélection. En effet, nous excluons Cask et Tiam1 pour le calcul des énergies de références. Pour cela, nous effectuons des recherches BLAST avec comme requête la séquence extraite du fichier PDB sur la base de données « Swiss-prot + trEmBL » d'Uniprot avec la matrice BLOSUM62, l'option « Gapped » et sans l'option « filtre ».

Nous obtenons un premier ensemble pour chaque cas en nous limitant aux homologues de bonne qualité au regard de la E-value et du pourcentage d'identité, tout en conservant en même temps une certaine diversité. Cela oblige pour certaines protéines à accepter des E-values plus hautes que 10^{-40} , notamment INAD et NHREF, respectivement 10^{-32} et 10^{-10} , pour avoir un nombre d'homologues suffisant. Ensuite, les redondances les plus flagrantes sont enlevées manuellement. Finalement, les ensembles se composent de 42 à 126 homologues, avec des pourcentages d'identité supérieurs à 66% excepté pour INAD où il a fallu descendre jusqu'à 38%, voir le tableau 5.4. Les alignements des séquences homologues retenues pour un groupe constitué des six premières protéines sont représentés aux figures 5.2, 5.4, 5.6, 5.8, 5.10 et 5.12.

5.4.4 Alignements des protéines expérimentales et leurs homologues

Afin d'obtenir une caractérisation structurale de notre sélection de protéines, nous réalisons un alignement des huit séquences natives, présentées en haut de la figure 5.1. Cet alignement nous sert de base pour la définition d'un alignement structural. Nous pouvons alors définir un cœur hydrophobe de nos protéines « PDZ », il est représenté au centre

Table 5.4 – Sélection des homologues.

protéines	nombre	E-value	% identité
NHREF	62	$\leq 10^{-32}$	67-95
INAD	42	$\leq 10^{-10}$	38-95
GRIP	48	$\leq 10^{-45}$	84-95
Syntenin	85	$\leq 10^{-43}$	85-95
DLG2	43	$\leq 10^{-41}$	78-95
PSD95	50	$\leq 10^{-46}$	81-95
Cask	126	$\leq 10^{-27}$	60-85
Tiam1	50	$\leq 10^{-22}$	60-85

de 5.1. Les 14 positions utilisées pour définir le cœur hydrophobe sont bien conservées dans l’alignement « seed » de Pfam, mais pas totalement. L’Arg, Lys et Gln apparaissent à certaines positions, puisque dans de petites protéines comme des domaines PDZ, la longue partie hydrophobe de ces chaînes latérales peut être enfouie dans le noyau tout en permettant à la pointe polaire de la chaîne d’être exposé au solvant. Quelques résidus Asp et Glu apparaissent aussi, dans les endroits où l’alignement des séquences peut ne pas très bien refléter la superposition 3D les chaînes latérales.

Table 5.5 – Similarité des séquences expérimentales homologues des huit protéines PDZ.

Protein	NHREF	INAD	GRIP	Syntenin	DLG2	PSD95	Cask	Tiam1
NHREF	326	64	15	15	59	112	49	1
INAD	64	221	56	-9	88	107	25	9
GRIP	15	56	378	24	65	87	90	39
Syntenin	15	-10	24	311	-26	22	42	-18
DLG2	59	88	65	-26	325	110	24	22
PSD95	112	107	87	22	110	325	66	21
Cask	49	25	90	42	23	66	308	37
Tiam1	1	10	39	-18	22	21	37	371

5.4.5 Similarité des homologues

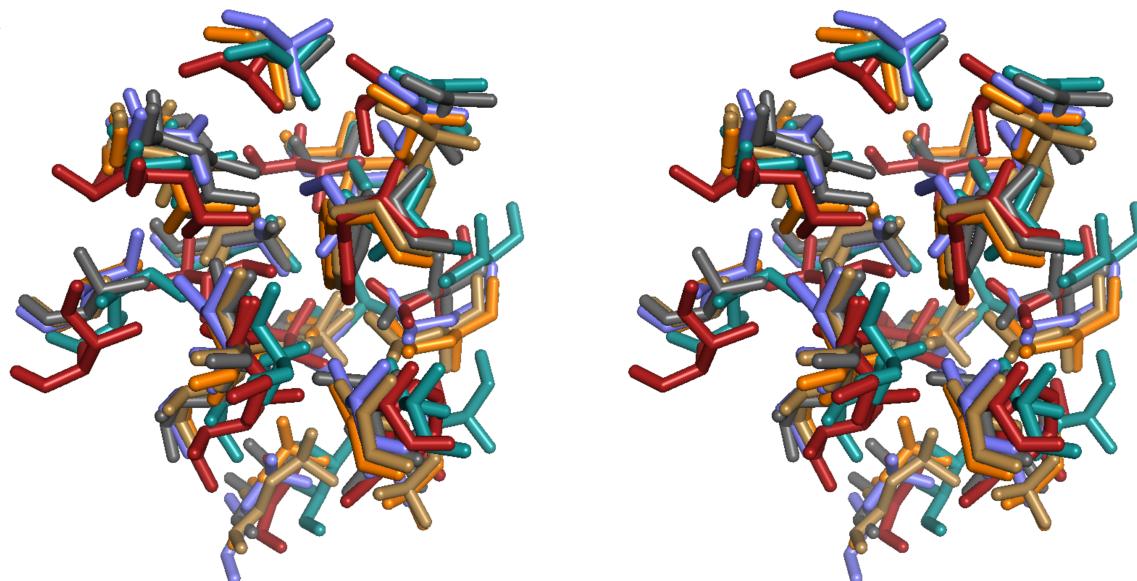
Comme nous avons caractérisé les homologies des séquences natives, nous calculons également la similarité des profils formés des séquences expérimentales homologues. Chaque couple de profils est aligné comme l’alignement du couple de séquences natives correspondant, voir figure 5.1. Apparaissent des situations où les similarités sont négatives, c’est le cas chez les homologues de Cask et Syntenine, protéines déjà repérées comme éloignées en

termes de séquences, avec une valeur de -26 entre DLG2 et Syntenine et -18 entre Tiam1 et Syntenin. Les plus hautes valeurs (excepté les valeurs d'auto similarité) sont 112 entre PSD95 et NREF et 110 entre PSD95 et DLG2. La similarité moyenne entre deux protéines différentes de notre sélection se situe aux environs de 50. Les détails sont à la table 5.5.

5.4.6 Les fréquences d'acides aminés

Pour chaque ensemble d'homologues, noté \mathcal{H} , nous calculons des fréquences globales de chaque type d'acide aminé sur toutes les séquences. Les fréquences sont déterminées séparément pour les positions enfouies et pour les positions exposées. Notons-les $f_t^b(\mathcal{H}), f_t^e(\mathcal{H})$, où l'indice t représente un type d'acide aminé et les exposants e et b désignent respectivement les ensembles de positions exposés et enfouis. Les ensembles de fréquences moyennes des huit protéines sont divisés selon deux groupes de protéines, d'une part le sous-ensemble $\mathcal{S}_1 = \{\text{NHREF, INAD, GRIP, Syntenin, DLG2, PSD95}\}$ et d'autre part le groupe $\mathcal{S}_2 = \{\text{Cask, Tiam1}\}$. Enfin, nous calculons la moyenne des $f_t^e(\mathcal{H})$ et des $f_t^b(\mathcal{H})$ sur \mathcal{S}_1 et sur \mathcal{S}_2 . Cela donne deux jeux de deux ensembles cibles de fréquences d'acides aminés $\{f_t^b\}$ et $\{f_t^e\}$. Nous faisons la même chose pour chaque classe de types. Cette partition en deux sous-ensembles va nous permettre d'estimer la transférabilité des énergies de références obtenues à partir d'un sous-ensemble de protéines sur l'autre.

1G90 RMLPRLCCLEK.GPN^YGFHLHGEGKGL.....GQY^IRLVEPGSPAEKAG.LLAGDRLVEVNGEN^YEKETHQQVS^RIRALA^NAVRLLVVDPETDEQL
 1IHJ GELIHMVTLDKTGKKSFGCIVRG^EVKDSPNTKTTGIFIKGIVPDSPAHLCGRLKGVDRLS^LNGKDVRNSTEQAVIDL^IKEADFKIELEIQT^F
 1N7E GAIYTVELKR.YGGPLGITISGTEEP.....FDPIISSLTKG^GLAERTGAIHIGDRILAINSS^LKGKPLSEAIHLLQMAGETVT^LKIKKQTD^AQPASS
 1R6J GAMDPRTITMHKDSTGHVG^FIFKN.....GKITSIVKDSSAARNG.LTTEHNICEINGONVIGLKDSQIADILSTSGTVV^TITIMPAF
 2BYG FQSM^TVVEIKLFK.GPKGLGFSIAGGVGNQH.IPGDNSIY^TTKI^DGGAAQKDGR^LQVGDRLLMVNNYSLEEVTHEEA^AV^ILKNTSEVVYLKV^GKPTT^I
 3K82 EDIPREP^RRIVIHR.GSTGLGFNIVGGE^E.....GIFISFILAGGPADLSGELRKGDQILSVNGVDLRNASHEQAAIALKNAGQT^VTIIAQYKPEEYSRF^A
 CASK RSRLVQFQKNTD^EPMG^ITLKM^NELN.....HCIVARIMHGGMIHRQGT^LHVG^DE^EREINGIS^VANQ^TV^EQLQKMLREM^RGSITFK^IV
 TIAM1 GAMGKVTHSIHIEKSDTA^DTYG^FSLSSVEED.....GIRR^LYVNSVKETGLASKKG.LKAGDE^ILEINNRAADALNSSMLKD^FL^SQP..SLGLLVR^TPEL



	Y	F	L	I	A	L	L	V	V	V	I	V	L	V
NHREF	24	26	28	39	48	53	59	62	67	75	79	86	88	90
INAD	F	I	I	I	A	L	I	L	V	V	I	I	L	I
GRIP	28	30	32	50	59	65	71	74	79	87	91	98	100	102
Syntenin	L	I	I	I	A	I	I	I	L	A	L	V	L	I
DLG2	682	684	686	698	707	713	719	722	727	735	739	746	748	750
PSD95	V	F	F	I	A	L	I	I	V	I	L	V	I	I
Cask	209	211	213	218	227	232	238	241	246	254	258	265	267	269
Tiam1	L	F	I	V	A	L	L	V	L	A	L	V	L	V
	323	325	327	338	347	353	359	362	367	375	379	386	388	390
	M	I	L	V	I	L	I	I	V	L	L	I	F	I
	501	503	505	515	524	530	536	539	544	552	556	563	565	567
	Y	F	L	V	A	L	I	I	A	L	L	L	L	V
	858	860	862	875	884	889	895	898	903	911	915	920	922	924

Figure 5.1 – Le cœur PDZ sélectionné En haut, l’alignement des huit séquences sauvages étudiées, les positions du cœur sont en jaunes. Au centre, la structure 3D des huit cœurs superposés. En bas, La séquence et les « positions PDB » de chaque cœur.

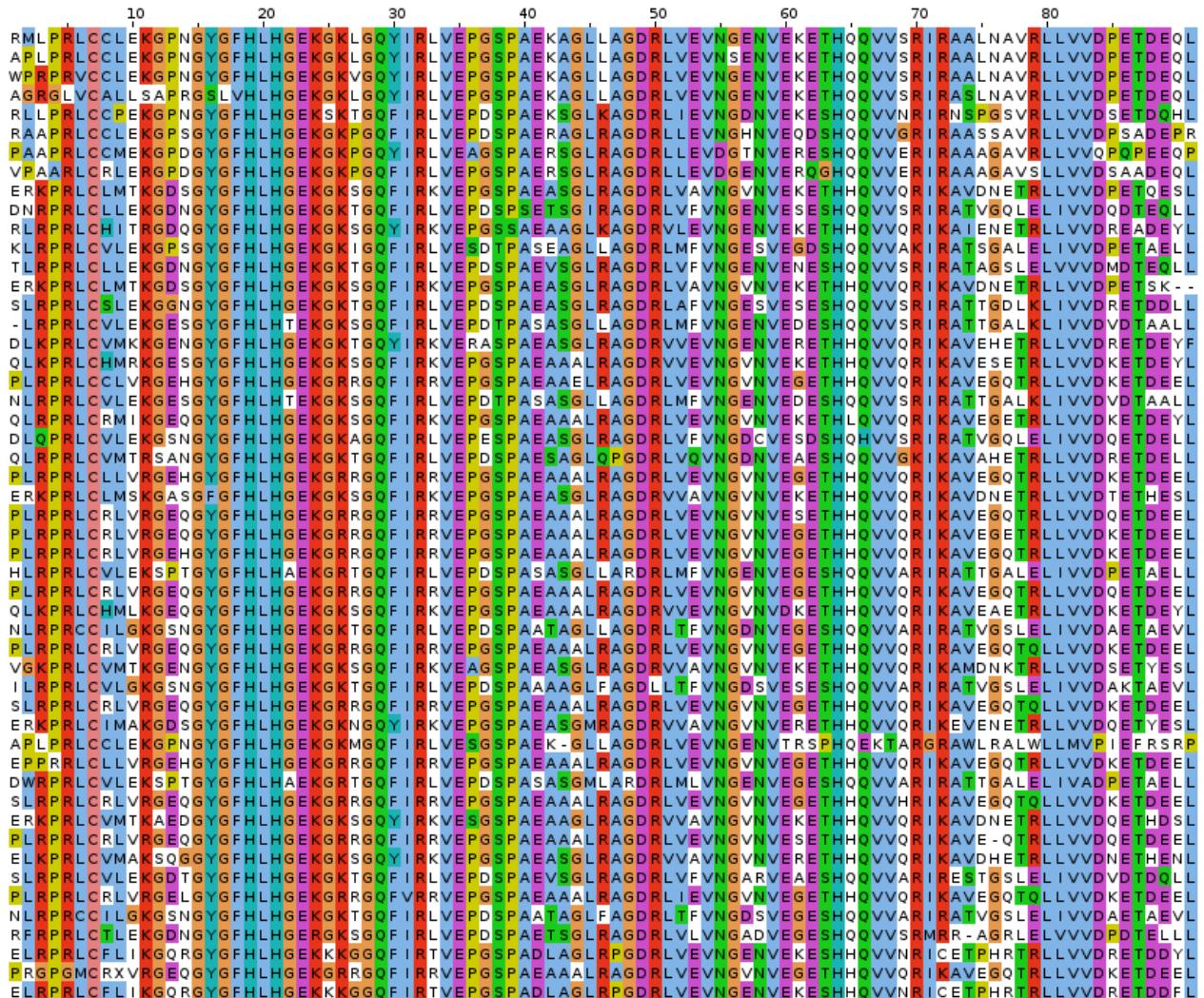


Figure 5.2 – L’alignement de notre sélection de séquences homologues à la protéine NHREF (code PDB : 1G9O)

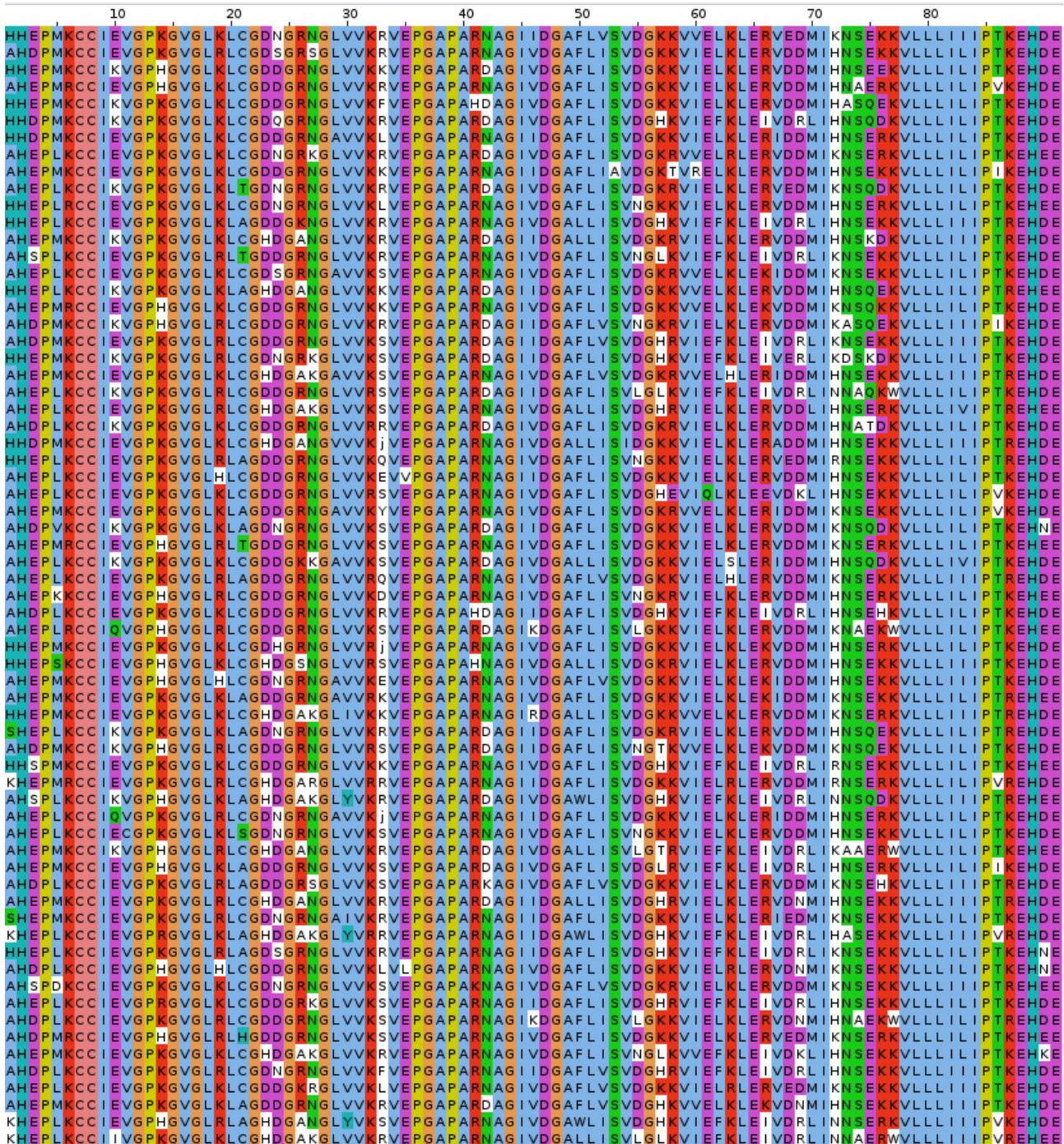


Figure 5.3 – Une sélection de séquences protéiques, parmi les 10 000 séquences de meilleure énergie, obtenues avec le backbone de la protéine NHREF (code PDB : 1G9O), modèle NEA

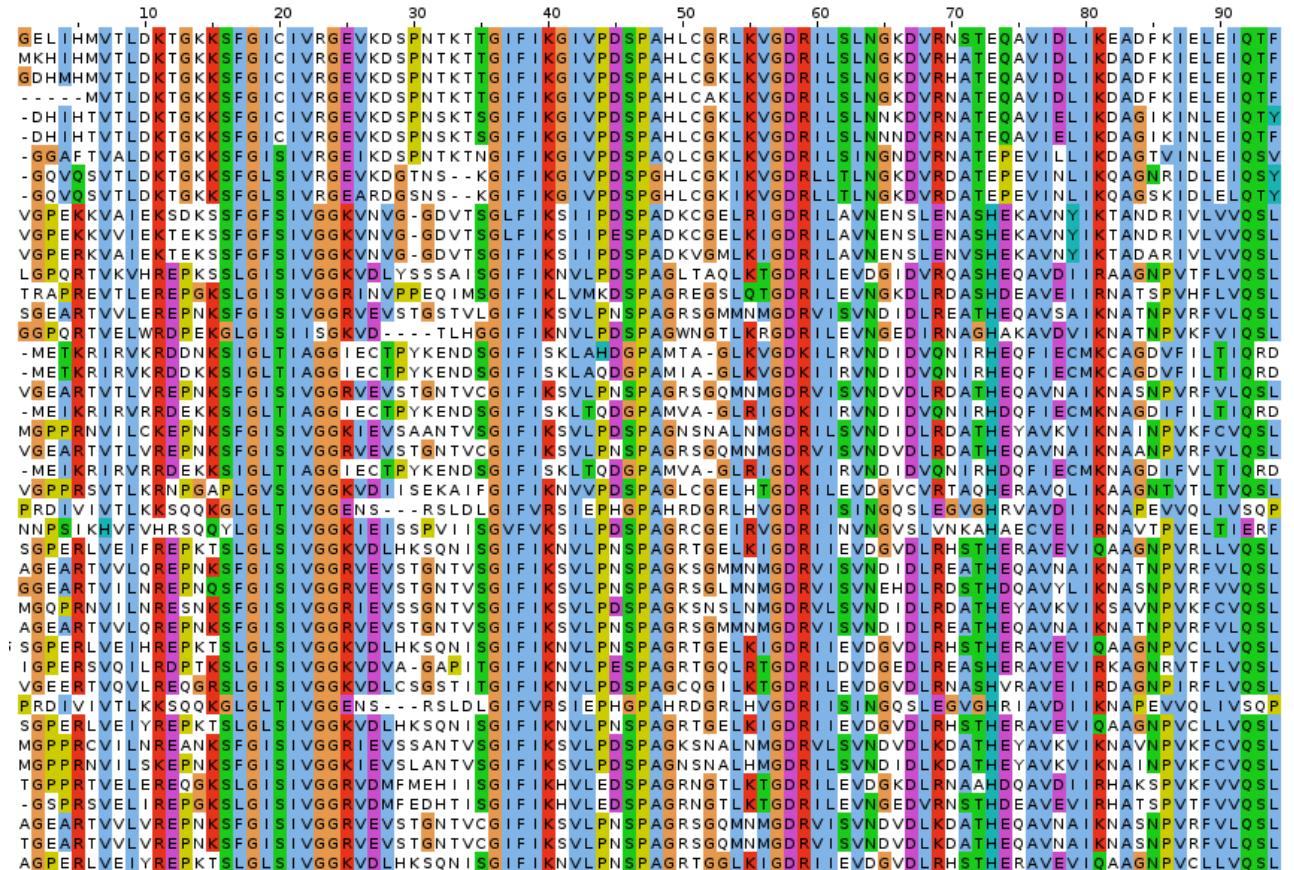


Figure 5.4 – L’alignement de notre sélection de séquences homologues à la protéine INAD (code PDB : 1IHJ)

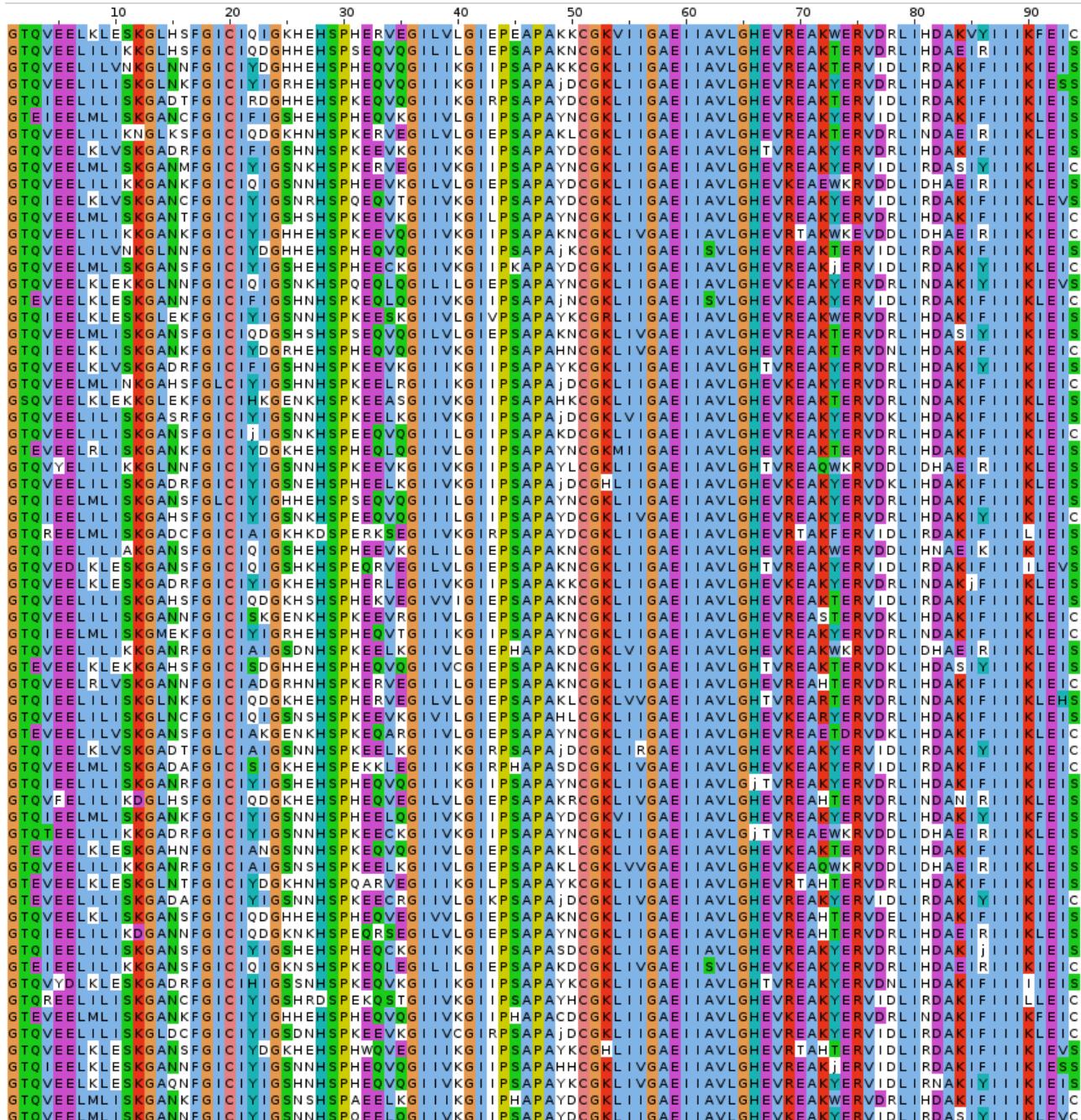


Figure 5.5 – Une sélection de séquences protéins, parmi les 10 000 séquences de meilleure énergie, obtenues avec le backbone de la protéine INAD (code PDB : 1IHJ), modèle NEA

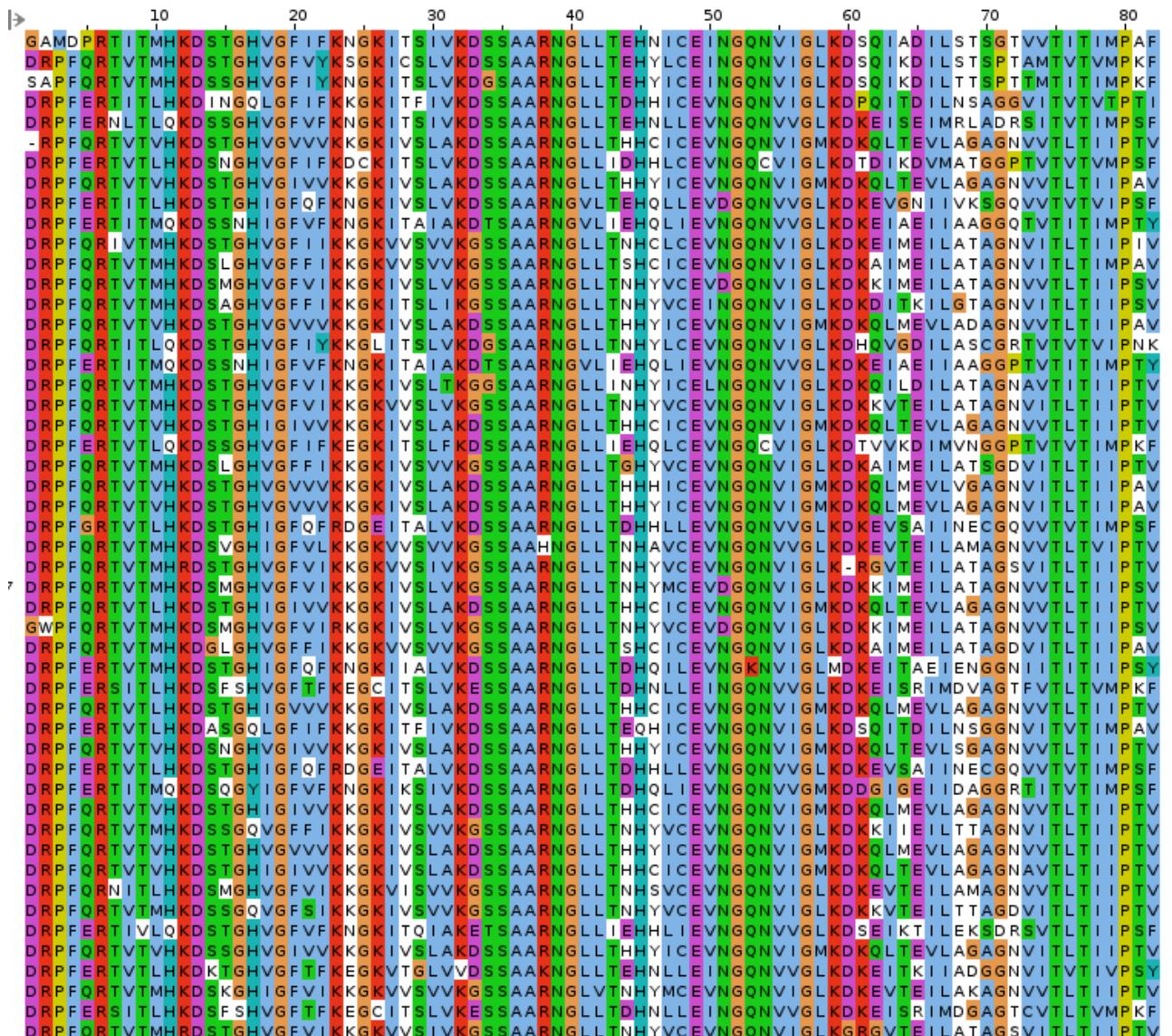


Figure 5.6 – L'alignement de notre sélection de séquences homologues à la protéine Syntenin (code PDB : 1R6J)

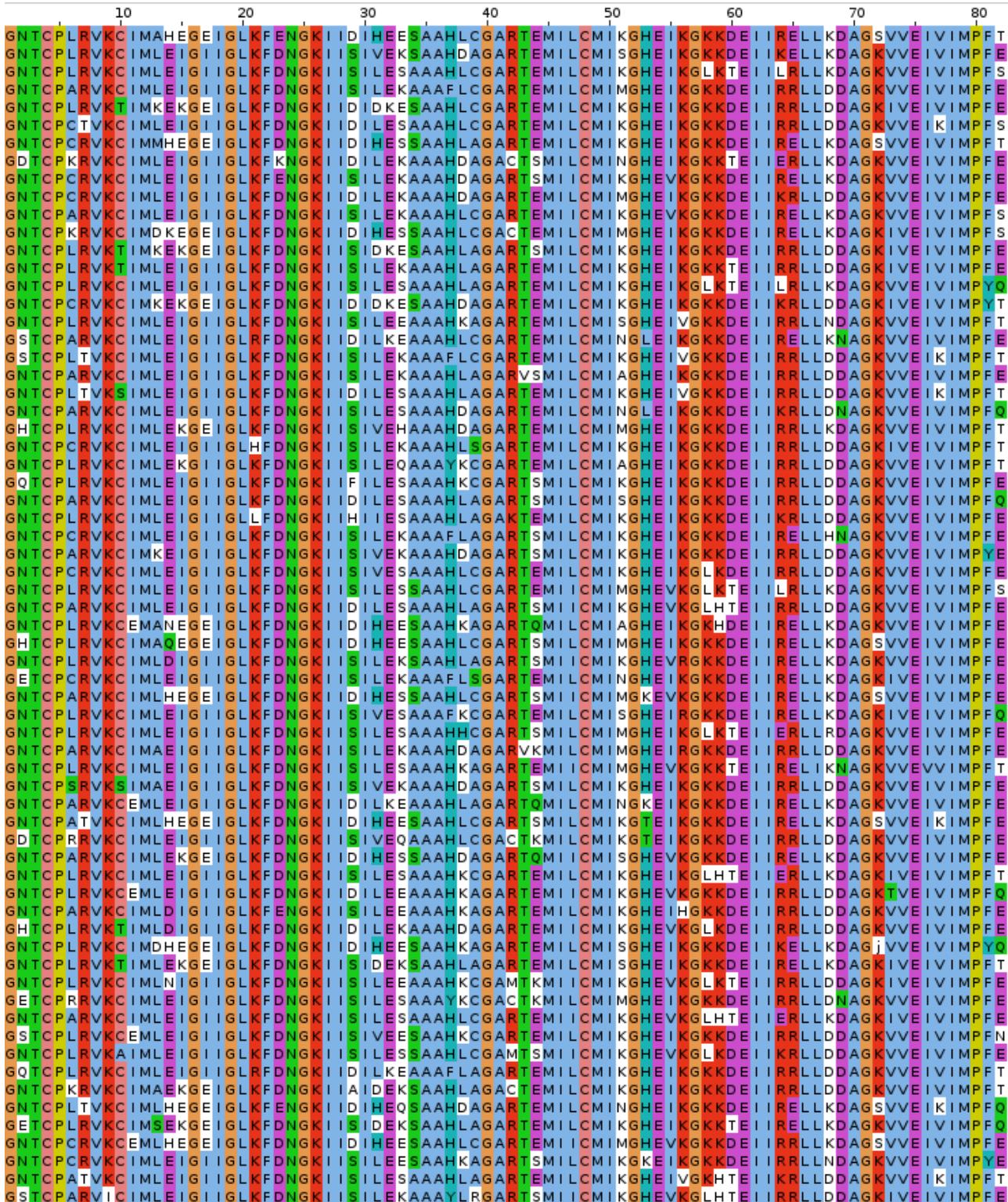


Figure 5.7 – Une sélection de séquences protéiques, parmi les 10 000 séquences de meilleure énergie, obtenues avec le backbone de la protéine Syntenin (code PDB : 1R6J), modèle NEA

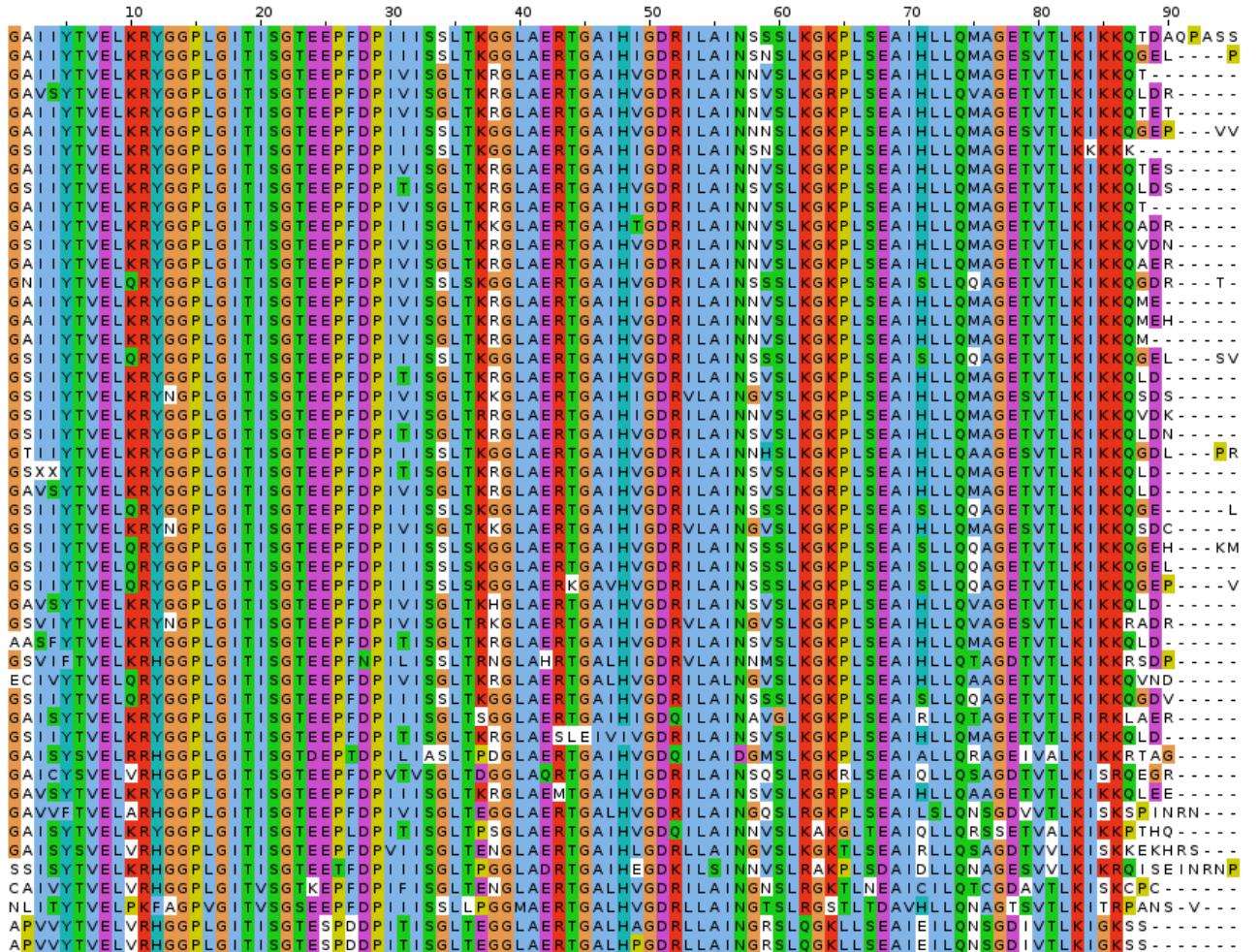


Figure 5.8 – L’alignement de notre sélection de séquences homologues à la protéine GRIP (code PDB : 1N7E)

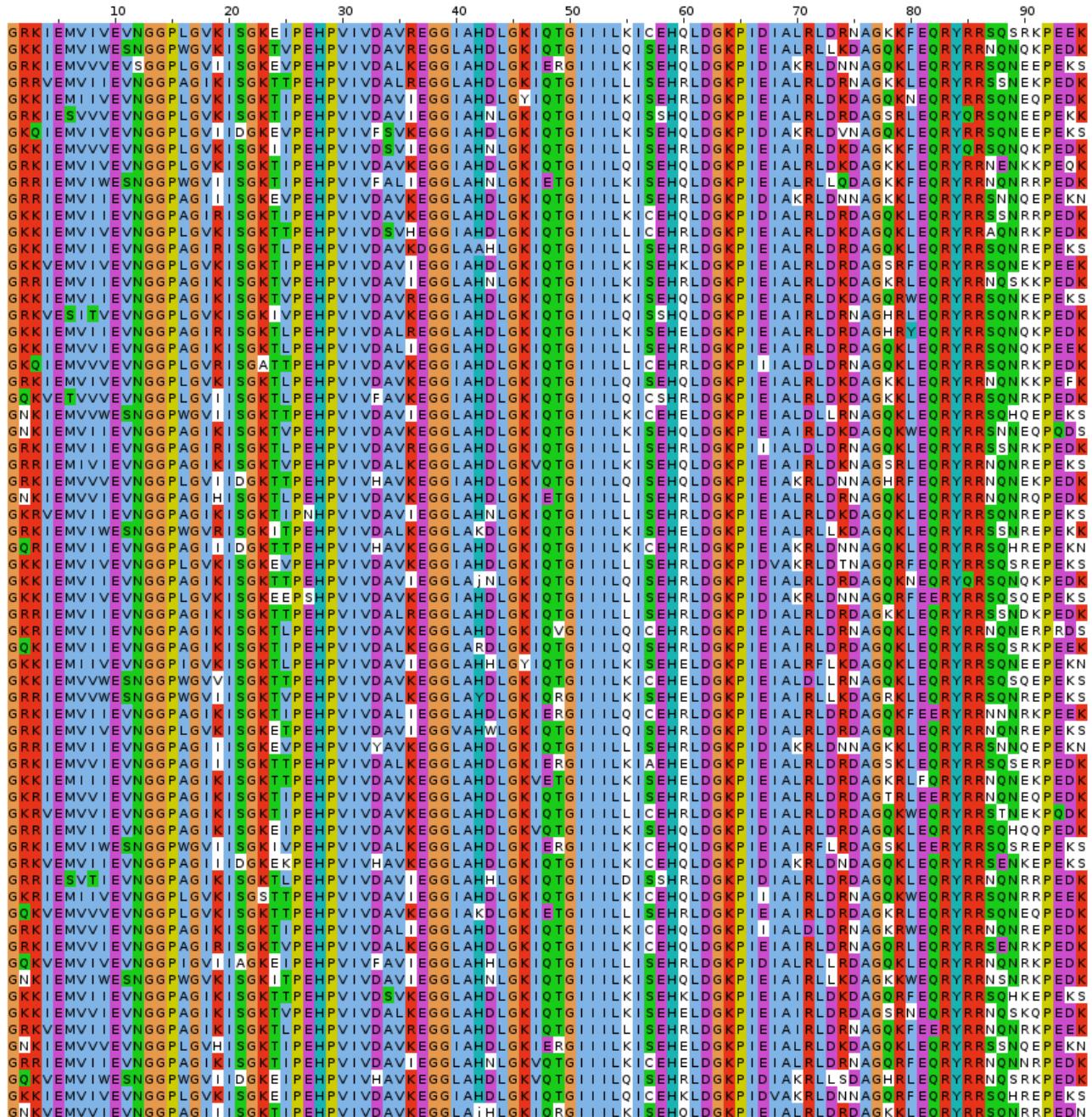


Figure 5.9 – Une sélection de séquences protéins, parmi les 10 000 séquences de meilleure énergie, obtenues avec le backbone de la protéine GRIP (code PDB : 1N7E), modèle NEA

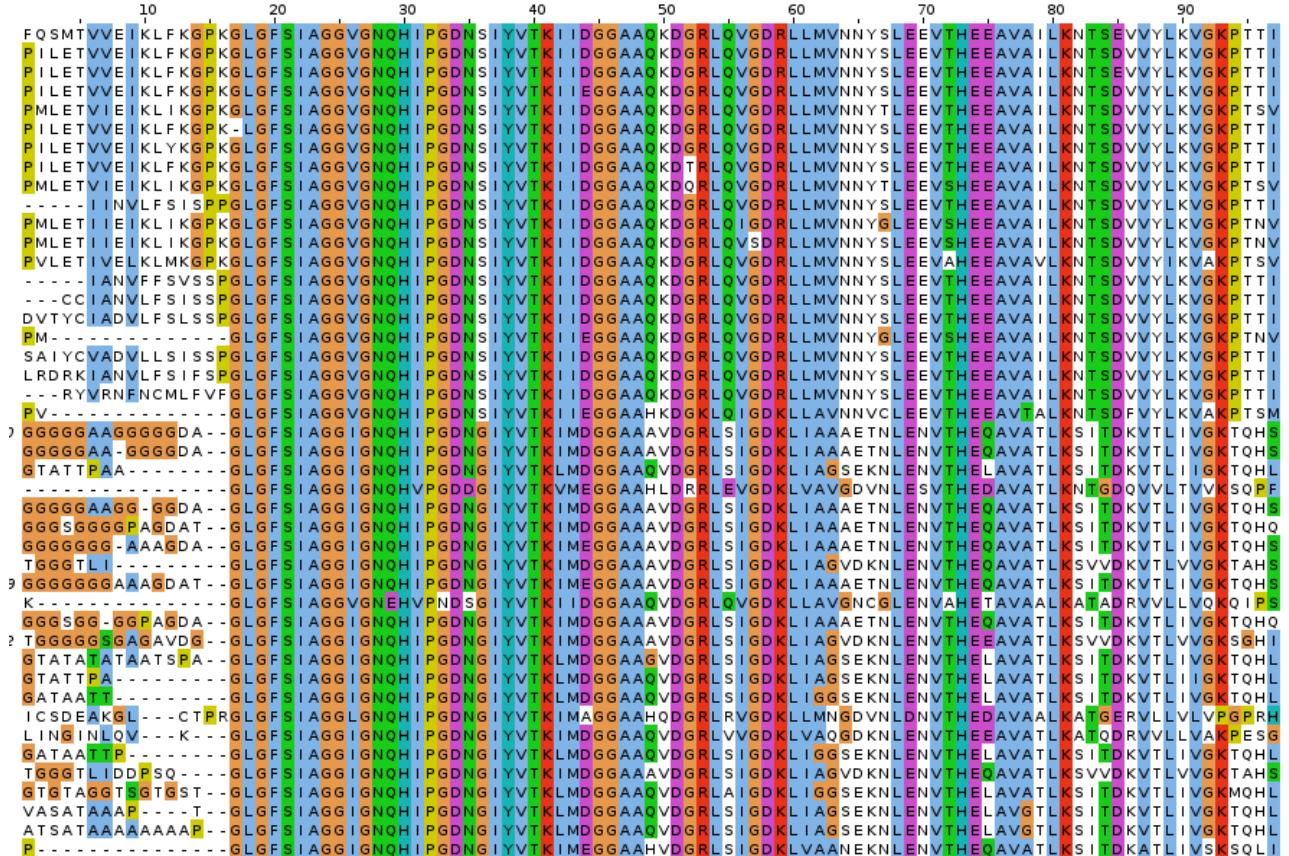


Figure 5.10 – L’alignement de notre sélection de séquences homologues à la protéine DLG2 (code PDB : 2BYG)

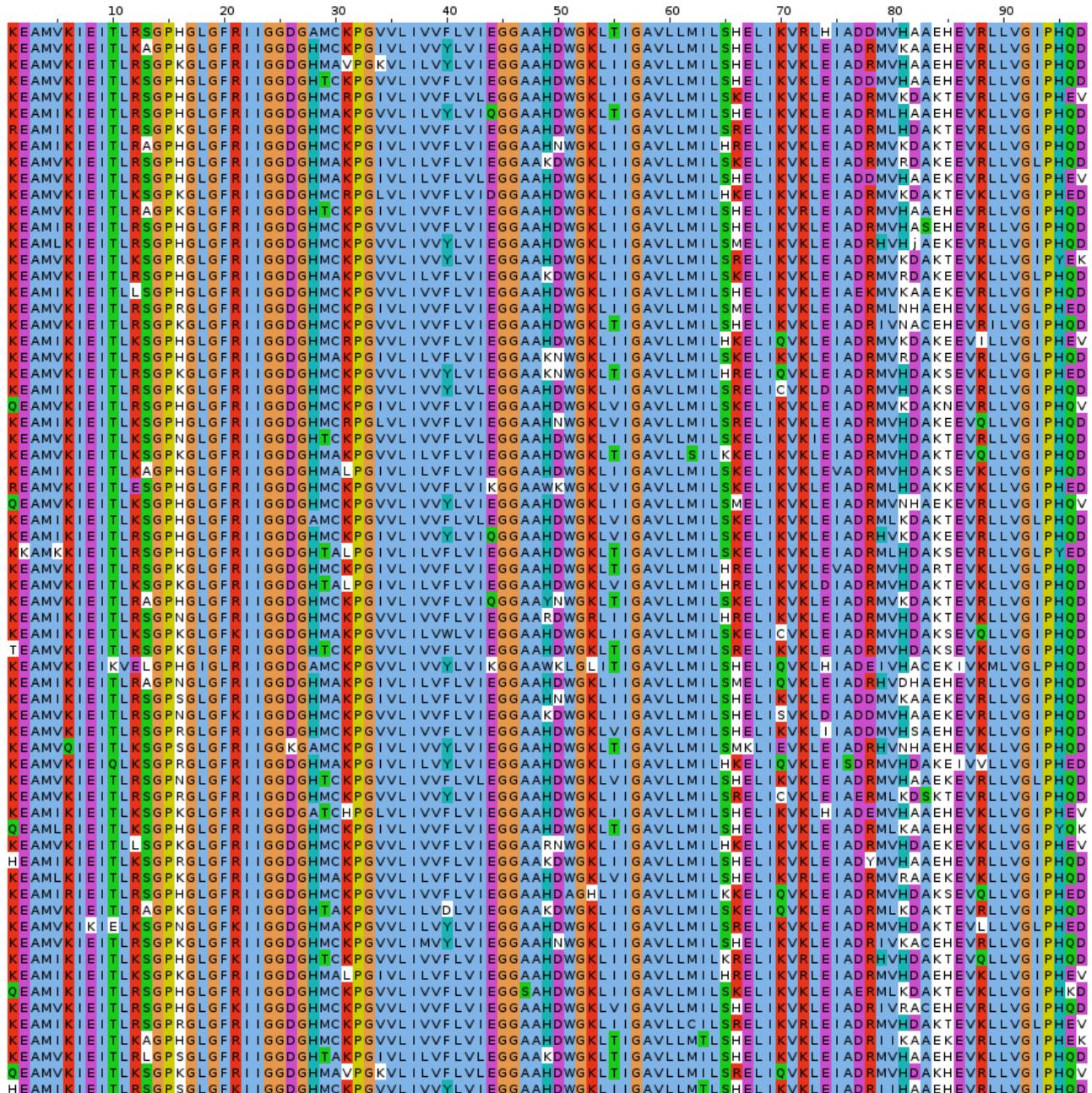


Figure 5.11 – Une sélection de séquences protéins, parmi les 10 000 séquences de meilleure énergie, obtenues avec le backbone de la protéine DLG2 (code PDB : 2BYG), modèle NEA

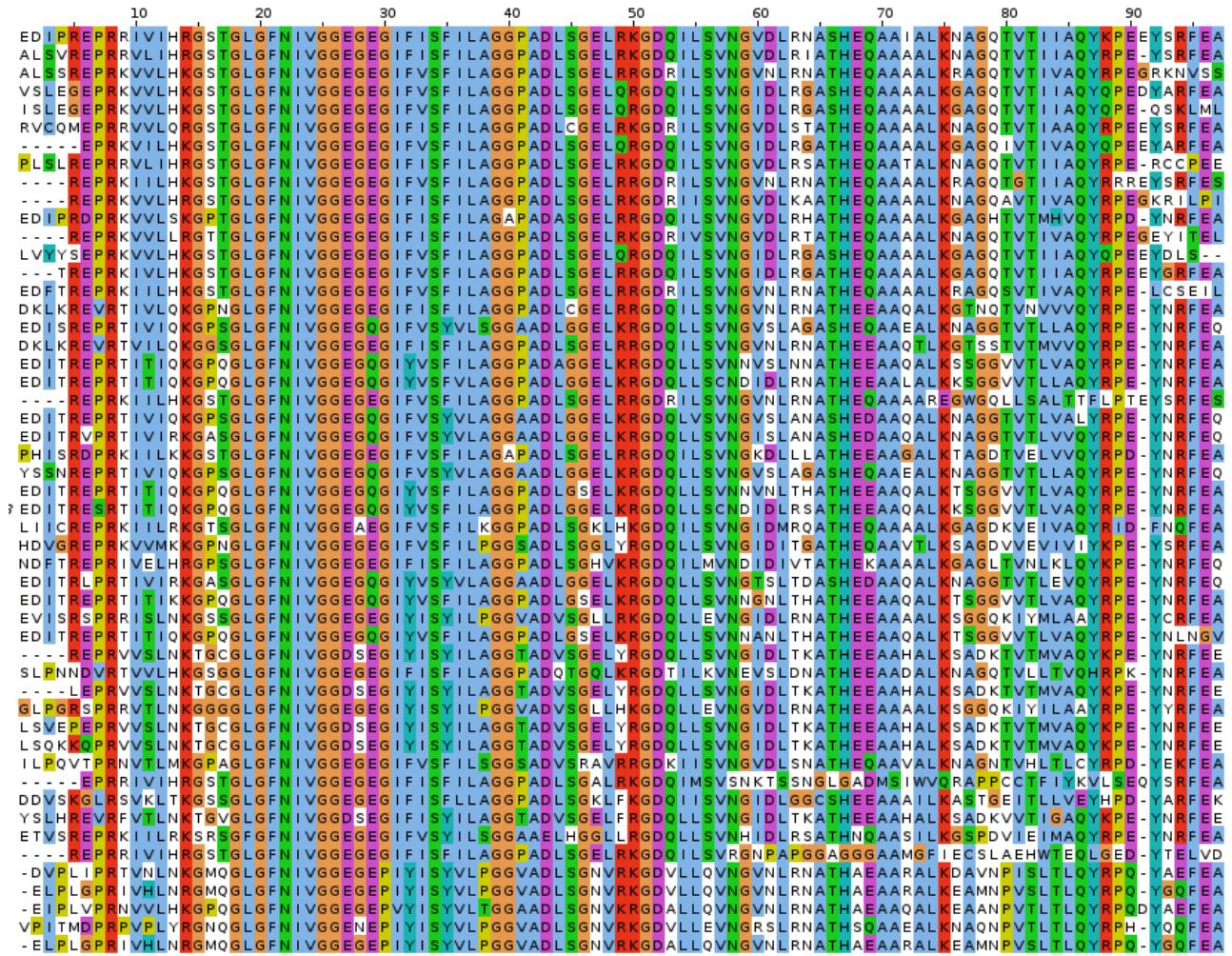


Figure 5.12 – L’alignement de notre sélection de séquences homologues à la protéine PSD95 (code PDB : 3K82)

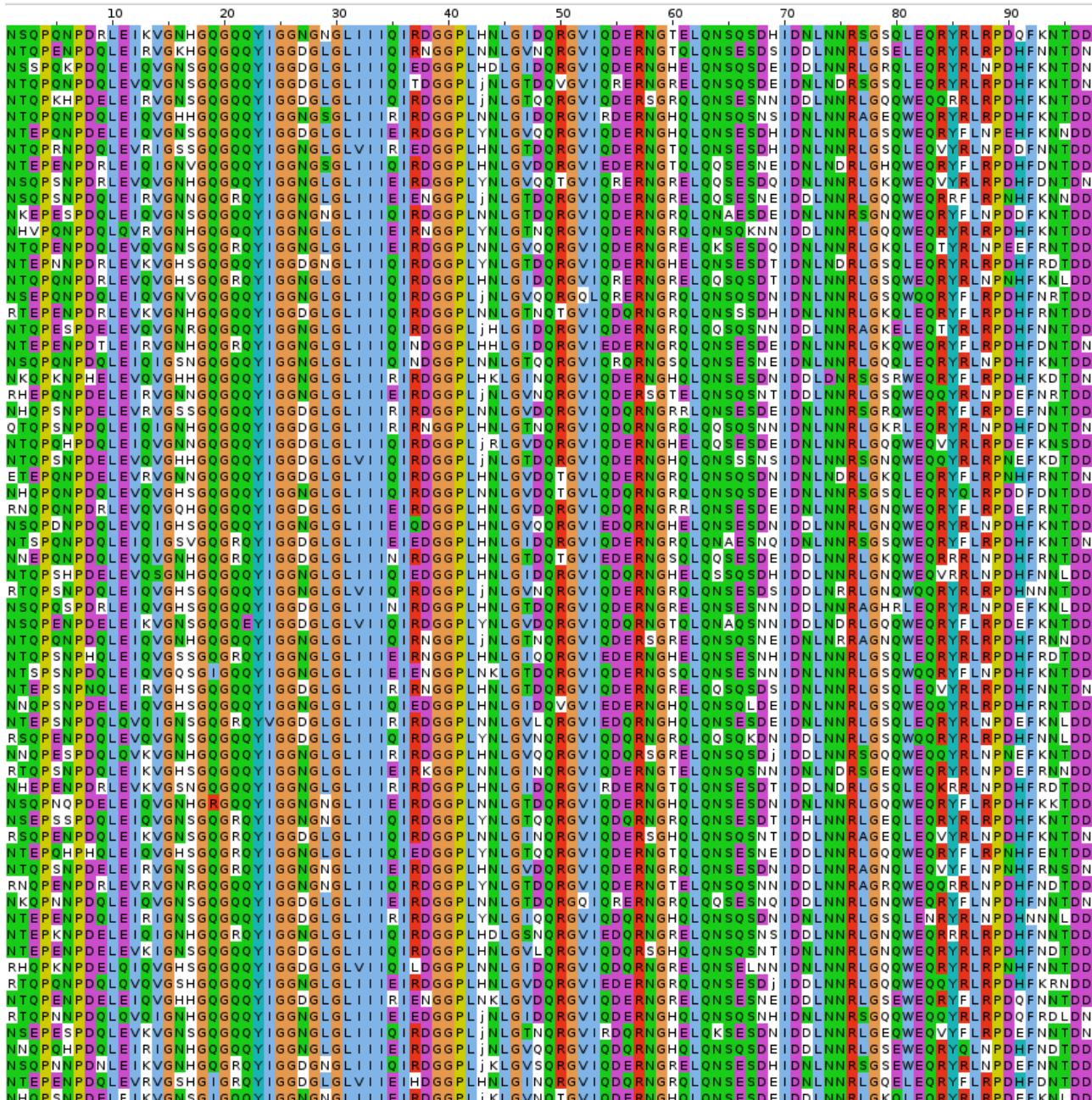


Figure 5.13 – Une sélection de séquences protéiques, parmi les 10 000 séquences de meilleure énergie, obtenues avec le backbone de la protéine PSD95 (code PDB : 3K82), modèle NEA

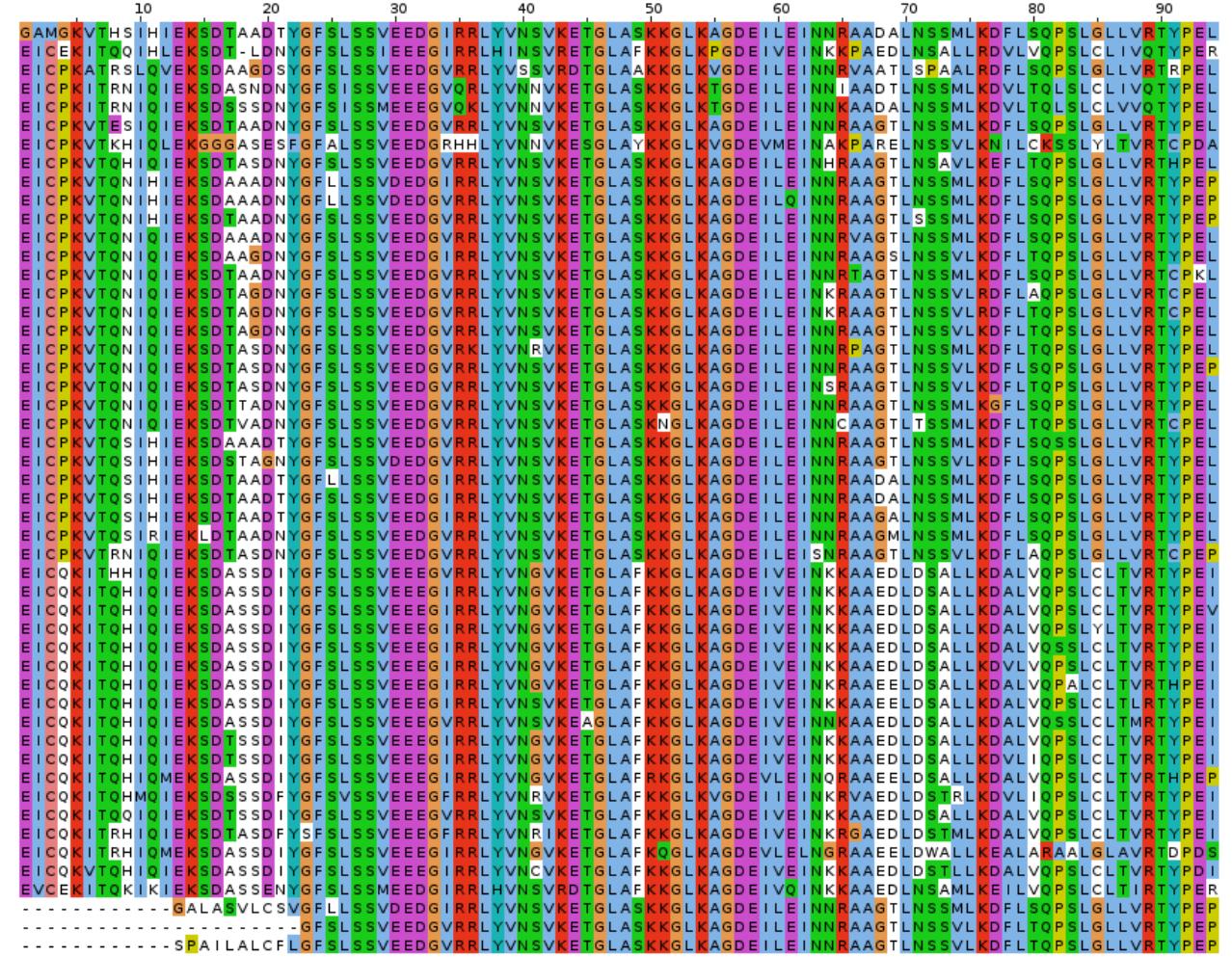


Figure 5.14 – L’alignement de notre sélection de séquences homologues à la protéine Tiam1 (code PDB)

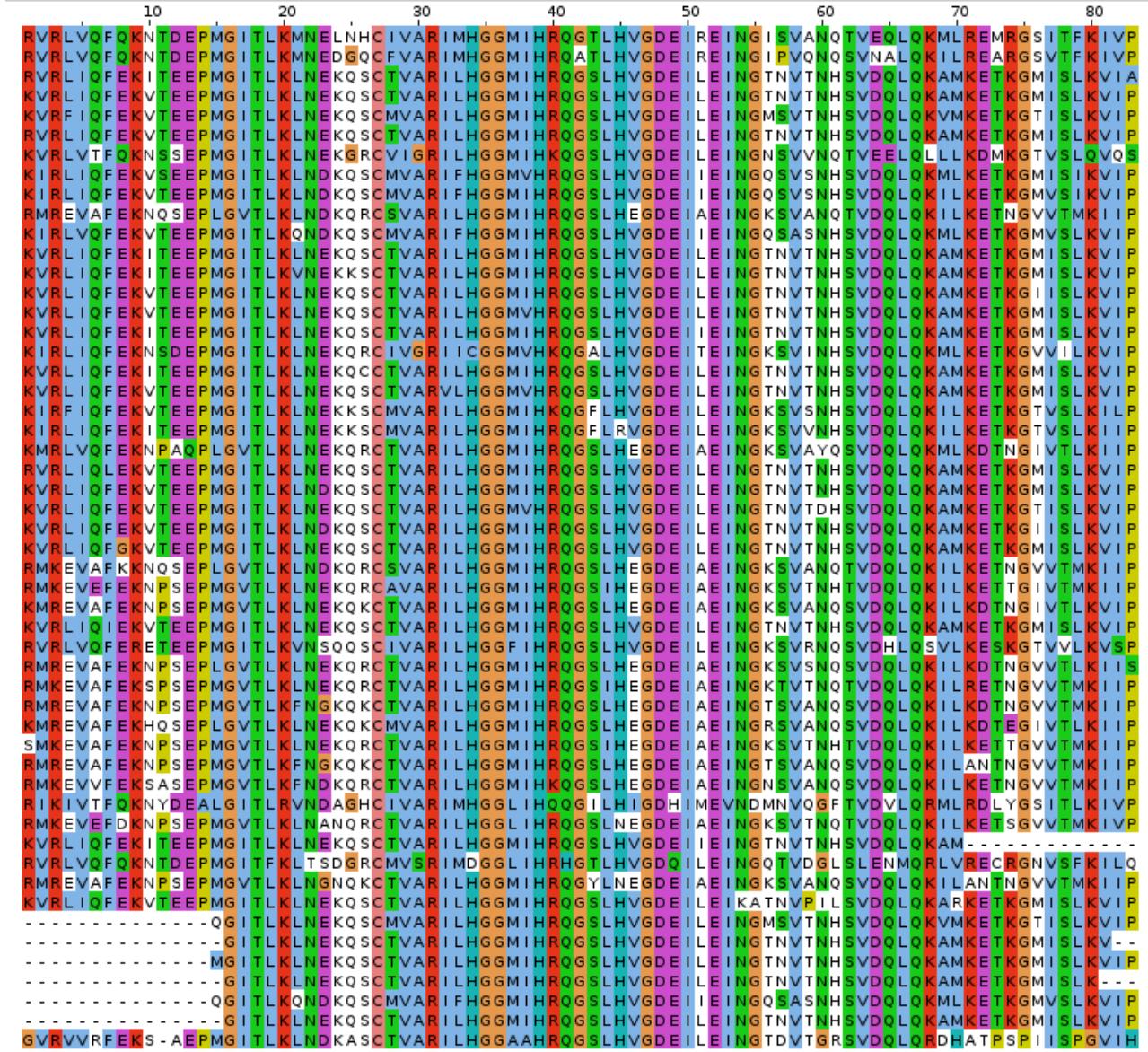


Figure 5.15 – L’alignement de notre sélection de séquences homologues à la protéine Cask (code PDB)

5.5 Séquences calculées

5.5.1 Préparation

Pour réaliser les calculs Monte Carlo, deux segments manquants dans le domaine Tiam1 (résidus 851-854 et 868-869) ont été construits en utilisant le programme Modeller [115]. Le ligand peptidique a été retiré de la structure PDB avant de calculer la matrice d'énergie. Les structures ont été préparées et les matrices d'énergie calculées à l'aide d'une procédure proposée dans des travaux précédents [90, 81] et détaillée en 2.4.2.

5.5.2 Simulation Monte Carlo

La production des séquences est réalisée avec Proteus [92]. Pour optimiser les énergies de références, les positions dans lesquels la séquence native comporte une glycine ou une proline conservent toujours leur type naturel et les positions mutables ne peuvent devenir ni Gly ni Pro. Pour optimiser les énergies de référence, nous sélectionnons, alternativement le long du backbone, une position pouvant muter, puis une position ne pouvant pas muter.

Dans tous les cas, des mutations sont produites au hasard, soumises uniquement à la fonction d'énergie MMGBSA qui entraîne la simulation. Les simulations Monte Carlo utilisent des mouvements à une ou deux positions, où les rotamères, les types d'acides aminés ou les deux peuvent changer. Pour les mouvements à deux positions, la deuxième position est choisie parmi celles qui ont une énergie d'interaction significative avec la première pour au moins une conformation du couple (10 kcal/mol ou plus). De plus, l'échantillonnage est amélioré par l'échange de répliques (REMC), où 8 simulations MC de 500 millions de pas sont exécutées en parallèle à différentes températures et avec échanges périodiques suivant le protocole REMCd qui est décrit en 4.2.2. Pour le calcul de fréquences, seules les séquences produites à la température $kT=0,263$ kcal/mol sont retenues.

5.5.3 Génération de séquence Rosetta

Des simulations Monte Carlo sont également réalisées à l'aide du programme et de la fonction d'énergie Rosetta [110]. Les simulations sont faites en utilisant la version 2015.38.58158 de la suite librement disponible en ligne, en utilisant la commande :

```
fixbb -s Cask.pdb -resfile Cask.res -nstruct 10000 -ex1 -ex2 -linmem_ig 10
```

où les options ex1 et ex2 activent une recherche améliorée des rotamères pour les chaînes latérales enfouies. La dernière option correspond au calcul de l'énergie « on the fly » au cours de la recherche MC, et les paramètres par défaut sont utilisés pour les autres

options. Comme pour les simulations Proteus, les résidus Gly et Pro présents dans la protéine sauvage ne sont pas autorisés à muter, et les positions qui mutent ne peuvent pas muter en Gly ou Pro. Des simulations sont exécutées pour chacun de nos domaines PDZ, jusqu'à obtenir 10 000 séquences uniques de faible énergie, ce qui correspond à des durées d'exécution d'environ 5 minutes par séquence sur un seul cœur d'un processeur Intel récent, pour un total de 10 heures par protéines en utilisant 80 cœurs. C'est comparable au coût des calculs Proteus, en comptant le temps de calcul de la matrice d'énergie plus celui des simulations Monte Carlo.

5.5.4 Caractérisation des séquences calculées

Les séquences calculées sont comparées à l'alignement Pfam pour la famille PDZ, en utilisant la matrice Blosum40 et une pénalité de gap de -6. Cette matrice est bien adaptée pour comparer des homologies éloignées (les séquences CPD et celles de Pfam). Chaque séquence Pfam est également comparée à l'alignement Pfam, ce qui permet de comparer des séquences calculées et un ensemble de domaines PDZ naturels. Pour ces comparaisons Pfam/Pfam, si un domaine test T fait partie de l'alignement, la comparaison T/T n'est pas prise en compte, pour être plus cohérent avec les comparaisons CPD/Pfam. L'alignement Pfam utilisée est le « RP55 » (voir la section 2.7.4). Les similitudes sont calculées pour les 14 résidus du cœur et pour l'ensemble des positions mutables de la protéine.

Les séquences calculées sont soumises à la bibliothèque de modèles de Markov Cachés Superfamily [94, 95] qui tente de classer les séquences selon la base de données structurelle SCOP [61], voir 2.7.1 pour les détails. Le programme hmmscan est exécuté avec une E-value seuil de 10^{-10} .

5.6 Résultats du modèle NEA

5.6.1 Optimisation du modèle de l'état déplié

Nous optimisons les énergies de référence E_t^r pour les six protéines de \mathcal{S}_1 , en utilisant leurs homologues naturels pour définir les fréquences d'acides aminés cibles. La constante diélectrique ϵ_p de la protéine est fixée à 8. La méthode d'optimisation est la méthode linéaire (voir 5.2.3). Nous effectuons 20 itérations avec la contrainte des classes et 20 itérations sans cette contrainte, l'optimisation se faisant alors sur tous les types possibles. La fonction proxy calculée sur les groupes de types converge autour des valeurs 0,06 et 0,07 (pour respectivement les énergies enfouies et exposées). Pour les types les valeurs sont 0,08 et 0,14. Cela correspond à des variations pour les E_t^r inférieures à 0,05 Kcal/mol

pour tous les types d'acides aminés sur les 5 derniers cycles d'optimisation. Le tableau 5.6 donne les énergies de référence convergées.

Table 5.6 – Les énergies de référence obtenues par l'optimisation sur six protéines.

Type d'acides aminés	enfouis	exposés
ALA	0,00	0,00
ARG	-28,29	-28,90
ASN	-5,94	-6,00
ASP	-9,19	-9,80
CYS	-1,04	-1,04
GLN	-4,72	-4,78
GLU	-7,90	-8,51
HID	11,96	12,39
HIE	11,43	11,85
HIP	14,53	14,96
ILE	4,72	2,11
LEU	1,17	-1,44
LYS	-4,56	-4,47
MET	-2,78	-3,54
PHE	-0,37	-2,55
SER	-3,73	-2,80
THR	-3,82	-3,82
TRP	-1,61	-3,79
TYR	-4,20	-6,10
VAL	0,83	-1,77

Le tableau 5.7 compare les fréquences d'acides aminés des homologues naturels et des séquences CPD. La population des différentes classes d'acides aminés a bien rejoint l'expérience, avec des écarts de moins de 1% dans la majorité des cas, pour les positions exposées et pour les positions enfouies. Seulement deux classes ont des écarts de plus de 2% (2,1 et 2,6 pour Lys et Arg aux positions exposées). L'accord pour les types d'acides aminés est également bon, avec seulement deux écarts de plus de 2% qui correspondent aux deux plus mauvaises classes. Pendant les 20 premiers cycles d'optimisation, la distribution des fréquences intra classe dépend par construction du δE_t^r défini dans pour chaque classe, qui sont calculés par mécanique moléculaire (voir section 5.2.1). La seconde série de 20 itérations permet l'ajustement de ces valeurs.

Table 5.7 – Les compositions en acide aminé (%) des séquences expérimentales et Proteus après optimisation des E_t^s . Les différences entre expérimentale et théorique sont indiquées entre parenthèses. Les types d'acides aminés sont rassemblés selon les groupes d'optimisations.

Res	Experimentale				Proteus			
	Enfoui	Exposé	Enfoui	Exposé				
ALA	10,9	4,6	11,1	4,4				
CYS	1,3	16,9	0,5	13,4	0,0	0,3	12,0	
THR	4,7		8,3		5,9	(0,1)	7,3	(-1,4)
ASP	4,3	6,8	6,0	17,9	4,5	6,7	5,6	16,7
GLU	2,5		11,9		2,2	(-0,1)	11,1	(-1,2)
ASN	2,6		6,7		2,5	4,7	7,5	14,0
GLN	2,1		5,5	12,2	2,2	(0,0)	6,5	(1,8)
HIP	1,2		5,0		1,0		5,2	
HIE	0,0	1,2	0,0	5,0	0,1	1,1	0,4	5,6
HID	0,0		0,0		0,0	(-0,1)	0,0	(0,6)
ILE	16,0		4,2		16,9		4,1	
VAL	16,5	50,7	5,4	14,0	16,7	52,1	5,6	14,0
LEU	18,2		4,4		18,5	(1,4)	4,3	(0,0)
LYS	2,5	2,5	10,9	10,9	1,5	1,5	13,0	
						(-1,0)		(2,1)
MET	0,9	0,9	1,5	1,5	1,6	1,6	1,4	
						(0,7)		(-0,1)
ARG	2,8	2,8	8,7	8,7	2,5	2,5	6,1	
						(-0,3)		(-2,6)
SER	5,3	5,3	7,6	7,6	4,3	4,3	8,7	
						(-1,0)		(1,1)
PHE	4,1		2,4		4,5	4,6	2,1	2,1
TRP	0,0	4,1	0,0	2,4	0,1	(0,5)	0,0	(-0,3)
TYR	2,6	2,6	1,2	1,2	2,2	2,2	0,4	0,4
						(-0,4)		(-0,8)
GLY	0,8		3,1		0,0	0,0	0,0	0,0
PRO	0,1	0,9	1,8	4,9	0,0	(-0,9)	0,0	(-4,9)

5.6.2 Tests de reconnaissance de famille

À partir de nos énergies optimisées, nous générerons des séquences pour chaque protéine. Les simulations Proteus utilisent l'algorithme REMC avec huit répliques (ou marcheurs), 750 millions de pas par réplique et des énergies thermiques kT qui varient de 0,125 à 3 kcal/mol. Il s'agit du protocole REMCd détaillé en 4.2.2. Toutes les positions sont autorisées à muter librement vers tous les types d'acides aminés exceptés Gly et Pro. Les simulations ont été faites avec la fonction d'énergie MMGBSA, sans aucun biais vers les séquences naturelles ni aucune limite sur le nombre de mutations. Les 10 000 séquences avec les énergies les plus faibles parmi celles échantillonnées par au moins un des marcheurs MC sont retenues pour l'analyse. De la même façon, 10 000 séquences produites par Rosetta ont été retenues. Ces séquences sont analysées par les outils de reconnaissance de pli « Superfamily » (voir 2.7.1). Avec une constante diélectrique de 8, nous avons obtenu un pourcentage élevé de séquences correctement associées à la famille et superfamille PDZ : 100% pour NHREF, INAD, GRIP et DLG2 , 99% pour Syntenin , seule PSD95 donne un score relativement mauvais de 47% pour la famille et 50% pour la superfamille. Les E-values sont inférieures à 10^{-3} pour les affectations à la famille. Ces valeurs sont semblables à celles obtenues par Rosetta pour les cinq premiers cas, par contre l'affectation à la superfamille est meilleure pour Rosetta avec des E-values compris entre $3,7 \cdot 10^{-23}$ et $1,3 \cdot 10^{-9}$, alors que Proteus obtient des E-values compris entre $4 \cdot 10^{-4}$ et $2 \cdot 10^{-12}$ si l'on exclut PSD95. Les détails sont présentés aux tables 5.8 et 5.9.

Table 5.8 – Résultats Superfamily pour les séquences Proteus avec le modèle NEA.

Protein	Match/seq size	Superfamily Evalue	Superfamily success	Family Evalue	Family success
NHREF	81/91	$2,00 \cdot 10^{-12}$	10000	$9,97 \cdot 10^{-3}$	10000
INAD	84/94	$4,80 \cdot 10^{-11}$	10000	$2,83 \cdot 10^{-3}$	10000
GRIP	82/95	$4,73 \cdot 10^{-8}$	10000	$5,56 \cdot 10^{-3}$	10000
Syntenin	63/91	$4,01 \cdot 10^{-4}$	9999	$1,05 \cdot 10^{-2}$	9999
DLG2	84/97	$3,82 \cdot 10^{-10}$	10000	$3,75 \cdot 10^{-3}$	10000
PSD95	46/97	$7,65 \cdot 10^{-1}$	5029	$4,06 \cdot 10^{-2}$	4719

5.6.3 Séquences et diversité de séquences

Une sélection des meilleures séquences calculées par Proteus, au sens de l'énergie, pour le sous-ensemble de six protéines est montrée aux figures 5.3, 5.5, 5.7, 5.9, 5.11 et 5.13. Les homologues naturels pour les huit protéines sont présentés aux figures 5.2, 5.4, 5.6, 5.8,

Table 5.9 – Résultats Superfamily pour les séquences Rosetta

Protein	Match/seq size	Superfamily Evalue	Superfamily success	Family Evalue	Family success
NHREF	79/91	$1,3 \cdot 10^{-13}$	10000	$2,2 \cdot 10^{-3}$	10000
INAD	85/94	$7,4 \cdot 10^{-14}$	10000	$3,7 \cdot 10^{-3}$	10000
GRIP	84/95	$2,2 \cdot 10^{-10}$	10000	$1,2 \cdot 10^{-3}$	10000
Syntenin	76/82	$7,3 \cdot 10^{-13}$	10 000	$1,8 \cdot 10^{-3}$	10 000
DLG2	86/97	$1,3 \cdot 10^{-9}$	10 000	$9,6 \cdot 10^{-4}$	10 000
PSD95	90/97	$3,7 \cdot 10^{-23}$	10 000	$5,2 \cdot 10^{-4}$	10 000

5.10, 5.12, 5.14 et 5.15. Comme dans de nombreuses études de CPD antérieures [116, 71], l'accord avec l'expérience pour les positions du cœur est très bon, alors que l'accord pour les résidus de surface est nettement plus faible.

La diversité des séquences naturelles et des séquences calculées est caractérisée l'entropie résiduelle (voir 2.7.7). Comme référence nous utilisons l'ensemble Pfam Seed qui est constitué d'un sous-ensemble représentatif des domaines PDZ naturels (2.7.4). L'entropie S_i est calculée aux positions i de chaque protéine. Nous caractérisons chaque protéine par la moyenne $\langle e^{S_i} \rangle$ sur les positions i . La diversité des séquences Rosetta est légèrement meilleure avec des valeurs comprises entre 1,40 et 1,68 alors que celles de Proteus sont comprises 1,24 et 1,55. La diversité de Pfam Seed correspond à une entropie (exponentielle) moyenne de 3,11. Le regroupement des séquences calculées de NHREF, INAD, GRIP, Syntenin, DLG2 et PSD95 donne une entropie de 2,88 avec Rosetta et 2,42 avec Proteus. Ainsi six géométries de backbone ne peuvent pas atteindre les mêmes niveaux de diversité que l'ensemble Seed, construit pour caractériser la diversité des domaines PDZ et notamment la diversité de leur squelette.

5.6.4 Scores de similarité Blosum

Les scores de similarité Blosum40 entre les séquences calculées et les séquences naturelles sur l'ensemble des positions sont globalement faibles (voir la figure 5.16). les similitudes Proteus et Rosetta chevauchent le bas du pic des scores naturels pour NHREF INAD, GRIP et DLG2 avec des valeurs Proteus en dessous de celles de Rosetta de quelques dizaines de points, environ 20 pour NHREF, mais près de 50 pour Syntenin. Les résultats Proteus pour PSD95 sont beaucoup moins bons que ceux de Rosetta avec un écart moyen de plus de 70. En ce qui concerne les résidus du cœur, montrés à la figure 5.17 la similitude des séquences calculées avec les séquences naturelles est beaucoup plus forte. Beaucoup de séquences Proteus ont des scores de plus de 30 pour NHREF, INAD et DLG2. Proteus

Table 5.10 – Moyenne de l'exponentielle de l'entropie sur les ensembles des positions des protéines

Protein	Proteus	Rosetta	Pfam seed
NHREF	1,38	1,45	3,15
INAD	1,37	1,55	3,06
GRIP	1,33	1,44	3,09
Syntenin	1,39	1,43	3,03
DLG2	1,24	1,57	3,11
PSD95	1,27	1,40	3,15
6prots	2,42	2,88	
CASK	1,55	1,68	3,15
TIAM1	1,22	1,57	3,15

fait jeu égal avec Rosetta sur NHREF et Syntenin, et est globalement meilleur sur INAD et DLG2, et moins bon sur GRIP et PSD95.

5.6.5 Tests de validation croisée

Cette partie a été effectuée en commun avec Nicolas Panel, doctorant de notre laboratoire. Les détails sont publiés dans [52]. Comme premier test de validation croisée, nous appliquons nos énergies de référence aux deux domaines de notre sous-ensemble \mathcal{S}_2 , qui ne fut pas utilisé dans la partie optimisation, voir 5.4.6. Nous générerons des séquences Tiam1 et Cask qui sont alors soumises aux tests Superfamily et aux calculs de similarité. La performance de Tiam1 sur la superfamille est 84,6%, un peu en dessous des protéines de \mathcal{S}_1 . Le score de Tiam1 pour la reconnaissance de la famille est de 76,6%. Les scores de Cask sont du même ordre que ceux de nos six premières protéines avec 100% de reconnaissance pour la famille et la superfamille, avec des E-values comparables.

Comme validation croisée supplémentaire, les énergies de références ont été optimisées par Nicolas Panel, en utilisant notre sous-ensemble \mathcal{S}_2 comme ensemble alternatif de domaines PDZ, et la troisième méthode d'optimisation, voir 5.12. Nous générerons alors, avec ce modèle, des séquences pour Syntenin et DLG2 qui font partie de \mathcal{S}_1 . Encore une fois, les scores sont proches de ceux obtenus avec le modèle optimisé sur \mathcal{S}_1 . Les reconnaissances sont à 100% et les E-values montent de $4,01 \cdot 10^{-4}$ à $1,3 \cdot 10^{-2}$ avec Syntenin pour la superfamille et de $3,82 \cdot 10^{-10}$ à $8,0 \cdot 10^{-9}$ pour DLG2.

Les histogrammes des scores de similarité Blosum montrent que les scores globaux pour Tiam1 et Cask sont très semblables pour les deux modèles. Pour DLG2 et Syntenine, nous calculons également les scores de similarité en utilisant les deux modèles. Les scores

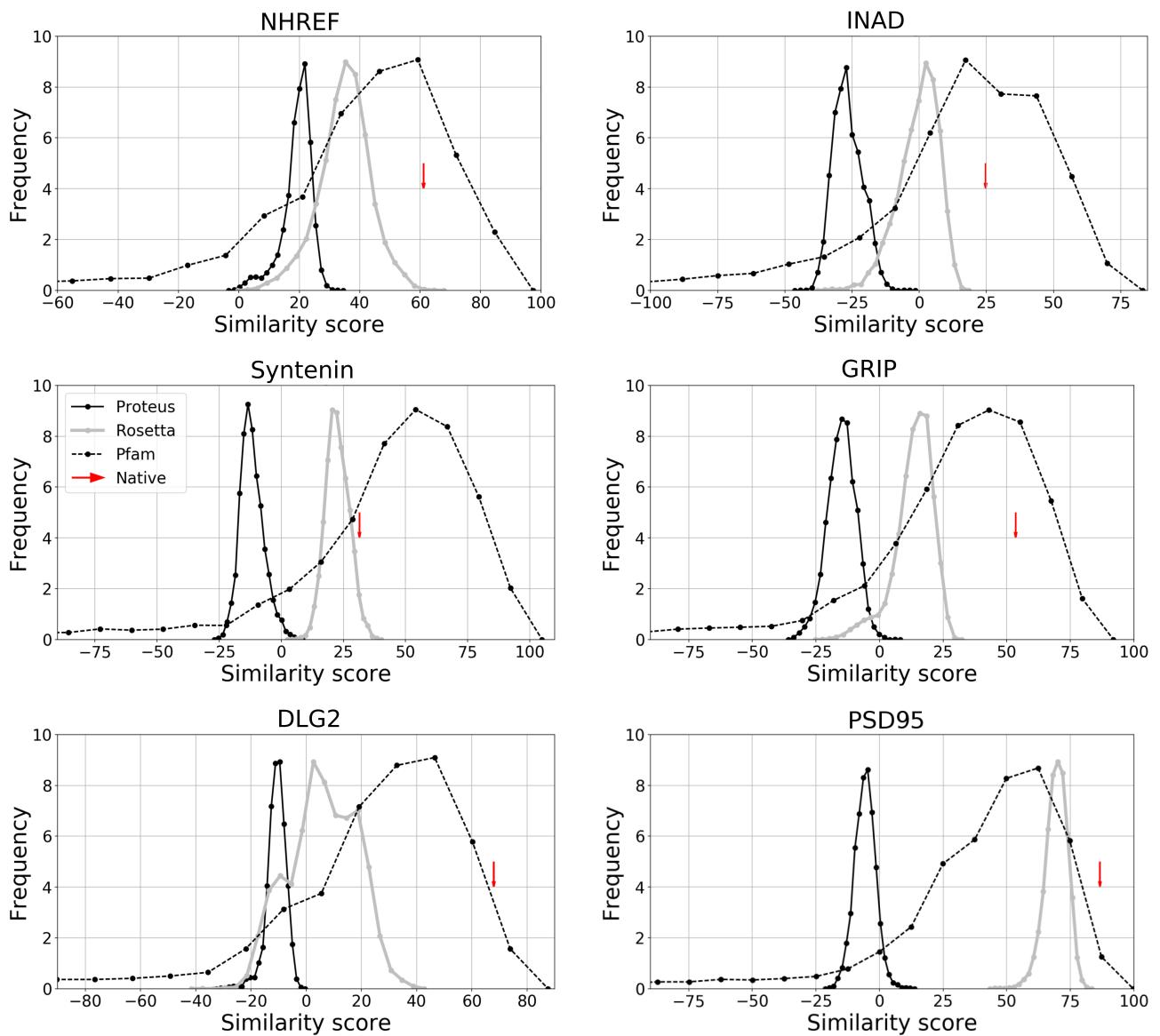


Figure 5.16 – Similarité des séquences des 6 protéines produites par Proteus et Rosetta à l’alignement Pfam RP55, sur l’ensemble des positions.

de similarité avec le modèle S_2 sont légèrement plus faibles qu’avec le modèle S_1 . Le score global a diminué d’environ 20 points pour la Synténine et environ 10 points pour DLG2. Dans l’ensemble, les modèles de validation croisée ont légèrement dégradé les performances. Ainsi, pour tout domaine d’intérêt PDZ, il semble préférable d’optimiser les énergies de référence spécifiquement pour ce domaine plutôt que de transférer des valeurs paramétrées en utilisant d’autres domaines PDZ.

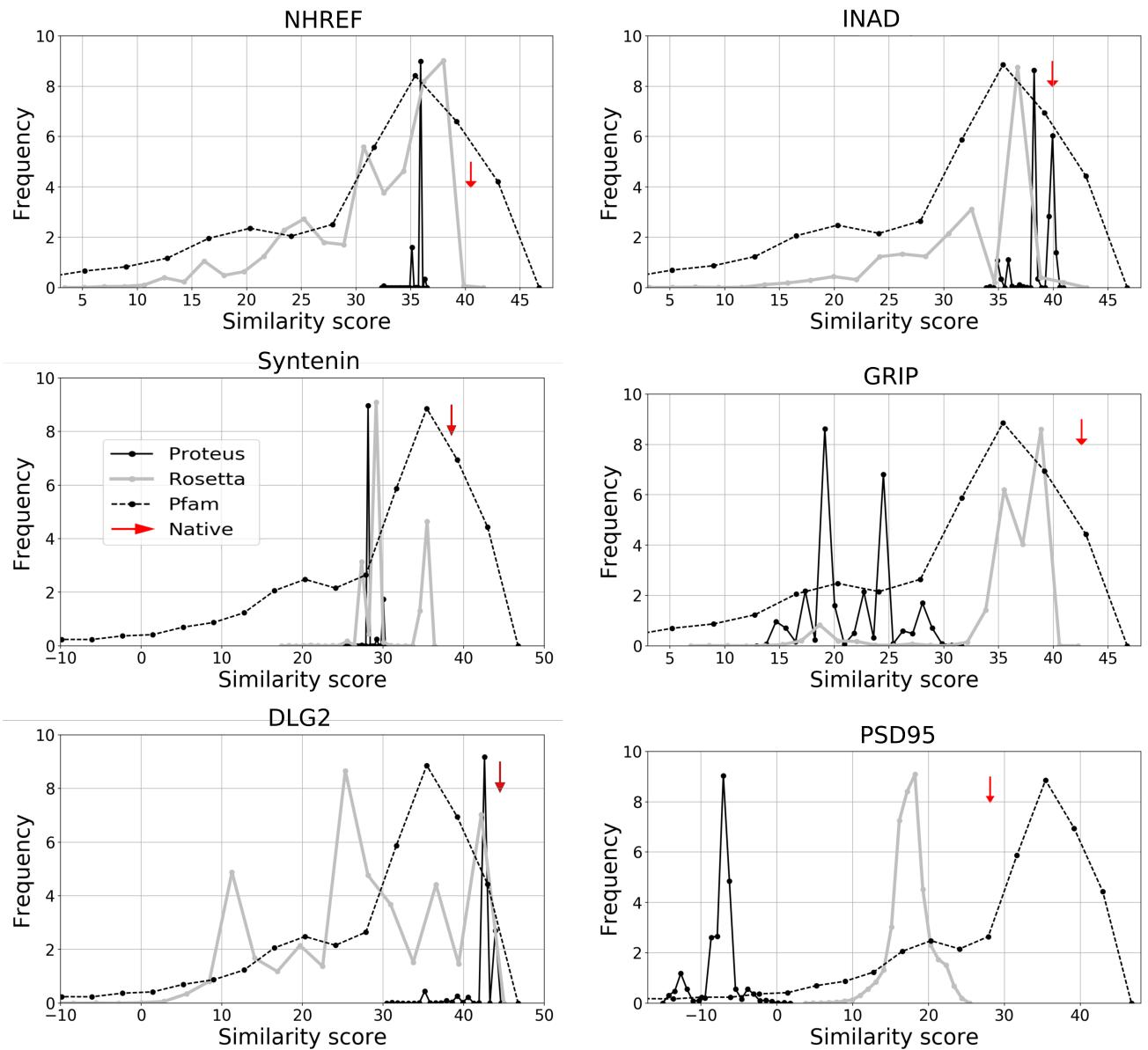


Figure 5.17 – Similarité des séquences des 6 protéines produites par proteus modèle NEA et Rosetta à l’alignement Pfam RP55, sur les positions du cœur hydrophobe.

5.7 Résultats du modèle FDB

5.7.1 Optimisation du modèle de l’état déplié

Nous optimisons maintenant les énergies de référence E_t^r en utilisant la variante FDB plus rigoureuse du modèle de solvant GB, voir le paragraphe 2.3.3. Avec ce variant, les fluctuations de la frontière protéine-solvant sont prises en compte (« Fluctuating Dielectric Boundary ») au prix d’une réorganisation du calcul. Nous nous limitons à trois protéines :

NHREF, Syntenin et DLG2. La constante diélectrique de la protéine est fixée à 4, et nous utilisons un jeu de paramètres surfaciques optimisés sur ces trois protéines (voir 5.3.1). La méthode d'optimisation est la méthode parabolique (voir 5.2.3) avec 20 itérations avec la contrainte des classes d'acides aminés et 20 itérations sans cette contrainte. Chaque itération se fait avec 100 millions de pas. La fonction proxy calculée sur les classes converge aux valeurs 0,06 et 0,03 pour les énergies enfouies et exposées et aux valeurs 0,03 (enfouis) et 0,02 (exposés) pour la fonction proxy calculées sur les types. Cela donne pour les E_t^r des fluctuations inférieures à 0,05 Kcal/mol sur les 5 derniers cycles. Pour tous les types sauf Arg et Hip dans le cas enfoui, où les variations montent à 0,1 Kcal/mol. La population des types d'acides aminés est proche de la population sauvage. Les écarts sur les classes d'acides aminés dans le cas des positions enfouies sont inférieurs à 1% sauf pour le groupe {Asp, Glu} (1,2%) et {Hip, Hie, Hid} (1,4%). Dans le cas exposé, les écarts sont un peu moins bons avec cinq groupes entre 2 et 3%. Les détails sont donnés dans le tableau 5.11.

Pour évaluer l'apport de l'optimisation FDB dans notre modèle, nous optimisons également les E_t^r pour nos trois protéines avec le modèle NEA et la constante diélectrique de la protéine à 4. La même méthode d'optimisation est utilisée. Dans ces conditions, les E_t^r se stabilisent à 0,05 kcal/mol près pour tous les types enfouis ou exposés. Ces quatre jeux d'énergies sont donnés dans le tableau 5.12. Ainsi nous avons deux modèles qui ne diffèrent plus que par la variante GB utilisée.

5.7.2 Tests de reconnaissance de famille

À présent, nous générions des séquences pour chaque protéine et pour chacun des jeux des E_t^r , FDB et NEA. Le protocole Proteus est identique à celui utilisé plus haut (section 5.6), la constante diélectrique de la protéine étant maintenant de 4. Ici encore, les 10 000 séquences de meilleures énergies parmi celles échantillonnées par les répliques REMC sont retenues pour l'analyse. Les résultats Superfamily sont présentés au tableau 5.13. Pour le FDB, la reconnaissance des familles et des superfamilles est de 100% pour les trois protéines tout comme Rosetta. En termes de E-value, Proteus FDB fait jeu égal avec Rosetta, avec des valeurs pour la superfamille allant de $2,85 \cdot 10^{-6}$ à $8,54 \cdot 10^{-14}$ pour Proteus et de $1,3 \cdot 10^{-9}$ à $1,3 \cdot 10^{-13}$ pour Rosetta et des valeurs très proches pour la famille. Pour le NEA, les résultats sont corrects pour la reconnaissance avec 98% pour les trois protéines, mais moins bons pour les E-values de la superfamille.

Table 5.11 – La composition en acide aminé (%) des séquences expérimentales et Proteus après optimisation FDB des énergies de référence. La différence entre expérimentales et Proteus sur les classes est donnée entre crochets.

Res	3 protéines expérimentales				FDB			
	Enfoui		Exposé		Enfoui		Exposé	
	type	classe	type	classe	type	classe	type	classe
ALA	8,7		5,5		9,6		1,8	
CYS	1,9	16,8	0,4	13,6	2,8	17,1 [-0,3]	0,6	11,3 [2,3]
THR	6,2		7,7		4,7		8,9	
SER	4,4	4,4	6,7	6,7	5,2	5,2 [-0,8]	7,9	7,9 [-1,2]
ASP	4,8		6,1		5,8	8,6	8,1	20,6
GLU	2,6	7,4	11,0	17,1	2,8	[-1,2]	12,5	[-3,5]
ASN	3,4		6,9		4,2	6,0	8,8	16,0
GLN	1,7	5,1	5,8	12,7	1,8	[-0,9]	7,2	[-3,3]
HIP	2,0		5,9		0,0		0,4	
HIE	0,0	2,0	0,0	5,9	0,5	0,6 [1,4]	2,7	5,0 [0,9]
HID	0,0		0,0		0,1		1,9	
ILE	12,4		4,1		11,2		0,6	
VAL	21,1	50,3	5,1	14,0	21,2	49,4 [0,9]	5,2	12,5 [1,5]
LEU	16,8		4,8		17,0		6,7	
MET	1,3	1,3	1,8	1,8	1,0	1,0 [0,3]	2,2	2,2 [-0,4]
LYS	3,4	3,4	10,8	10,8	4,2	4,2 [-0,8]	12,4	12,4 [-1,6]
ARG	1,1	1,1	9,8	9,8	1,8	1,8 [-0,7]	11,8	11,8 [-2,0]
PHE	4,6		2,4		3,9	3,9	0,0	0,0
TRP	0,0	4,6	0,1	2,5	0,0	[0,7]	0,0	[2,5]
TYR	2,4	2,4	1,2	1,2	2,0	2,0 [0,4]	0,0	0,0 [1,2]
GLY	1,0		1,9		0,0	0,0	0,0	0,0
PRO	0,1	1,1	1,8	3,7	0,0	0,0 [1,1]	0,0	[3,7]

Table 5.12 – Les énergies de référence obtenues avec l'optimisation sur 3 protéines. La constante diélectrique est fixée à 4.

acides aminés	NEA		FDB	
	Pos.	Enf.	Pos	Exp.
ALA	0,00	0,00	0,00	0,00
CYS	-0,89	-2,57	-1,06	-1,64
THR	-5,31	-8,075	-4,84	-6,68
SER	-5,55	-6,55	-4,45	-5,24
ASP	-17,26	-22,06	-14,56	-18,82
GLU	-16,12	-20,68	-14,52	-18,21
ASN	-16,38	-20,41	-14,02	-17,80
GLN	-14,00	-18,41	-13,14	-16,61
HID	11,21	6,95	10,85	8,13
HIE	10,63	6,15	10,41	7,37
HIP	15,17	10,72	12,86	10,98
ARG	-53,40	-57,36	-51,37	-54,76
LYS	-8,20	-12,34	-8,24	-11,35
ILE	6,76	3,44	5,50	3,06
VAL	0,43	-2,19	-0,05	-1,66
LEU	0,52	-3,72	0,00	-2,94
MET	-1,61	-3,21	-2,85	-3,09
PHE	1,86	-2,68	0,17	-3,18
TRP	-0,23	-7,67	-1,94	-5,53
TYR	-5,10	-10,90	-5,91	-10,14

5.7.3 Scores de similarité Blosum

Les scores de similarité Blosum40 sont calculés entre les séquences CPD et les séquences Pfam. Nous calculons ces similarités pour les séquences Proteus FDB, Proteus NEA et Rosetta, sur l'ensemble des positions sauf une petite partie au début et une petite partie à la fin de la séquence, parce qu'elles ne sont pas conservées dans l'alignement Pfam. Cela représente moins de 10% de la longueur de la séquence. Pour NHREF, l'approximation FDB améliore très nettement les scores de Proteus NEA avec un gain d'environ 50 points. Cela place Proteus à peu près au niveau de Rosetta. Pour Syntenin, les écarts sont plus serrés, avec le FDB légèrement moins bon que le NEA, mais proche de Rosetta. Dans le cas de DLG2 le FDB domine les séquences NEA et près de la moitié de celles produites par Rosetta. Sur les positions du cœur hydrophobe, les résultats Proteus sont excellents, avec le plus souvent des similarités avec Pfam compris entre 30 et 40. Le NEA et le FDB font jeu égal sur DLG2, le NEA étant au-dessus pour les deux autres. Il n'y a donc pas

Model	Protein size	Match/seq E-value	Superfamily success	Superfamily E-value	Family success	Family
Proteus FDB epsilon=4	NHREF	80/91	8,54 10 ⁻¹⁴	10000	8,94 10 ⁻³	10000
	Syntenin	70/82	2,85 10 ⁻⁶	10000	2,69 10 ⁻³	10000
	DLG2	88/97	3,26 10 ⁻¹²	10000	1,96 10 ⁻³	10000
Rosetta	NHREF	79/91	1,3 10 ⁻¹³	10000	2,2 10 ⁻³	10000
	Syntenin	76/82	7,3 10 ⁻¹³	10000	1,8 10 ⁻³	10000
	DLG2	86/97	1,3 10 ⁻⁹	10 000	9,6 10 ⁻⁴	10 000
Proteus NEA epsilon=4	NHREF	62/91	3,22 10 ⁻³	9857	1,00 10 ⁻²	9857
	Syntenin	70/82	2,83 10 ⁻³	9879	3,62 10 ⁻³	9879
	DLG2	83/97	1,66 10 ⁻³	9876	3,18 10 ⁻³	9876

Table 5.13 – Résultats Superfamily pour les séquences Proteus avec le modèle FDB (énergies de références optimisées sur 20 cycles selon les classes + 20 cycles selon les types).

d'amélioration sur le cœur ce qui s'explique par le fait que ces résidus sont trop éloignés du solvant pour bénéficier de FDB. Les scores Rosetta pour Syntenin sont proches, mais pour les deux autres protéines, les scores sont nettement plus variables et globalement moins bons. Tout cela est représenté à la figure 5.18.

5.7.4 Taux d'identité à la séquence native

Pour chaque protéine, nous calculons le taux d'identité des 10 000 séquences de meilleures énergies par rapport à la séquence native, ainsi que pour les 10 000 séquences Rosetta, voir 2.7.3. On considère uniquement les positions mutables sans celles aux extrémités des séquences comme expliqué au paragraphe précédent. Ce taux varie pour Proteus FDB entre 24% et 33% et entre 20% et 33% pour la version NEA. Pour Rosetta les taux se situent entre 35 et 40% avec un écart de 11% pour NHREF sur le FDB et de 7% pour les deux autres protéines (voir la table 5.14). Il s'avère donc que les séquences Rosetta sont nettement plus proches des séquences natives que celles de Proteus, avec sept mutations de moins en moyenne.

Séquences	Proteus FDB	Proteus NEA	Rosetta
NHREF	24	20	35
Syntenin	31	33	38
DLG2	33	31	40

Table 5.14 – Pourcentage d'identité moyen à la séquence native

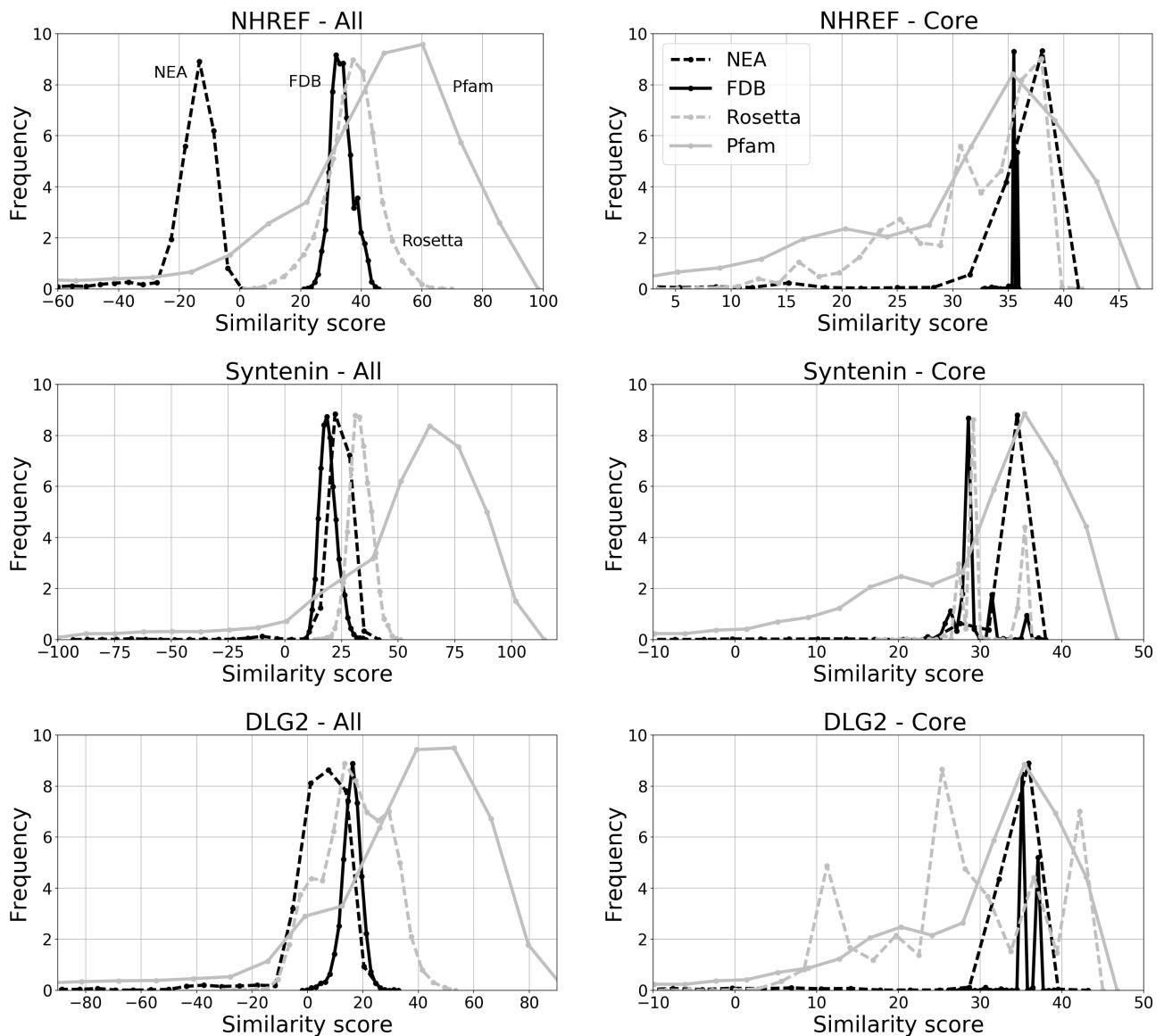


Figure 5.18 – Similarité des séquences Proteus (NEA et FDB), Rosetta et des séquences de l’alignement Pfam RP55, sur toutes les positions à gauche et sur les positions du cœur à droite.

5.7.5 Logos des séquences obtenues

Finalement, nous montrons les séquences obtenues. Elles sont représentées sous forme de logos à la figure 5.19 pour les positions du cœur hydrophobe, et la figure 5.20 pour les positions exposées. L’accord avec les séquences naturelles sur les positions du cœur est très bon et l’accord sur les positions exposées est nettement moins bon. Mais au regard de la diversité des types aux positions exposées dans les séquences naturelles de Pfam, le consensus entre les séquences naturelles est lui-même quasi inexistant.

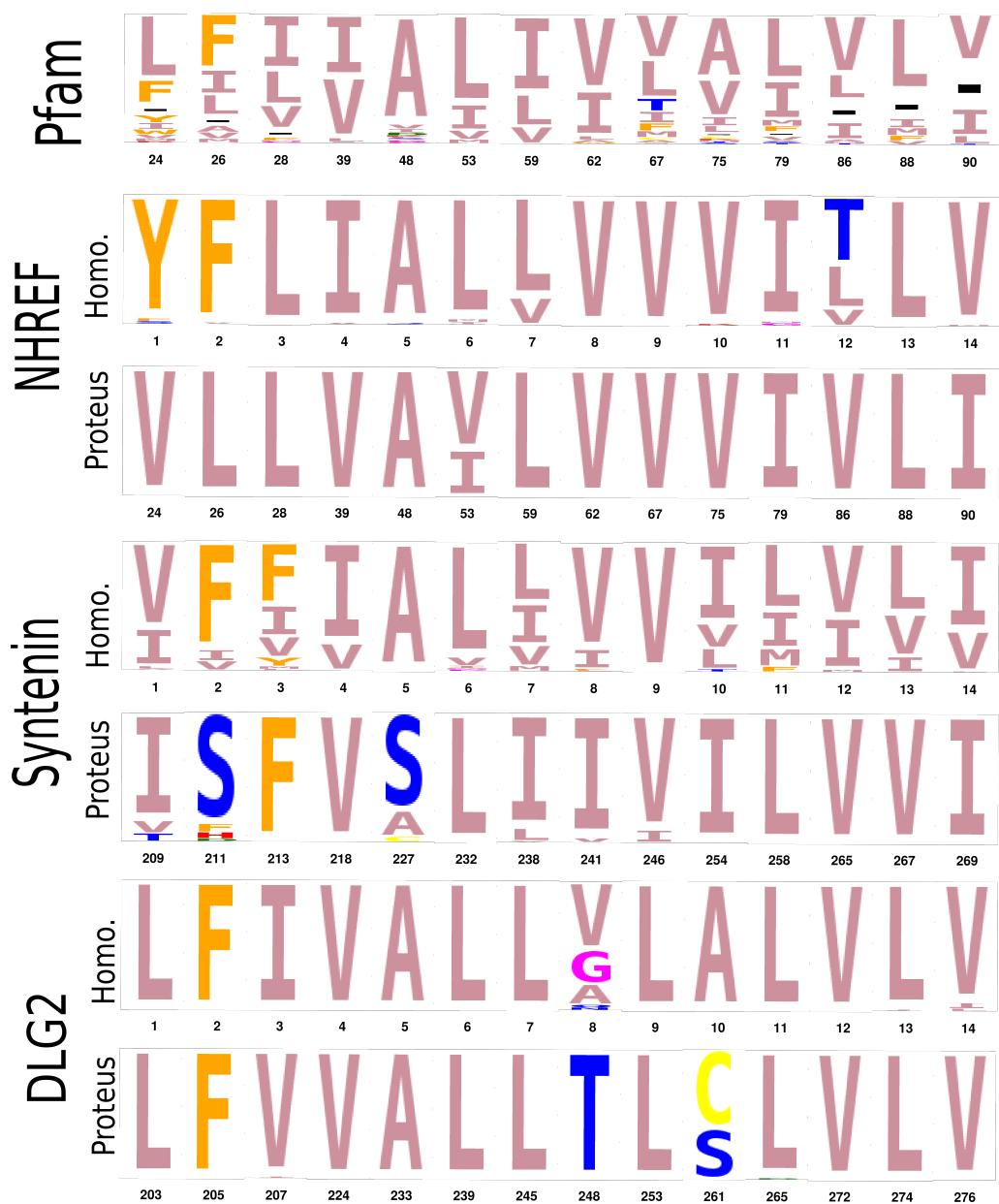


Figure 5.19 – Les séquences conçues par Proteus, les homologues à la native et les séquences naturelles représentées sous forme de logo pour les positions du cœur hydrophobe

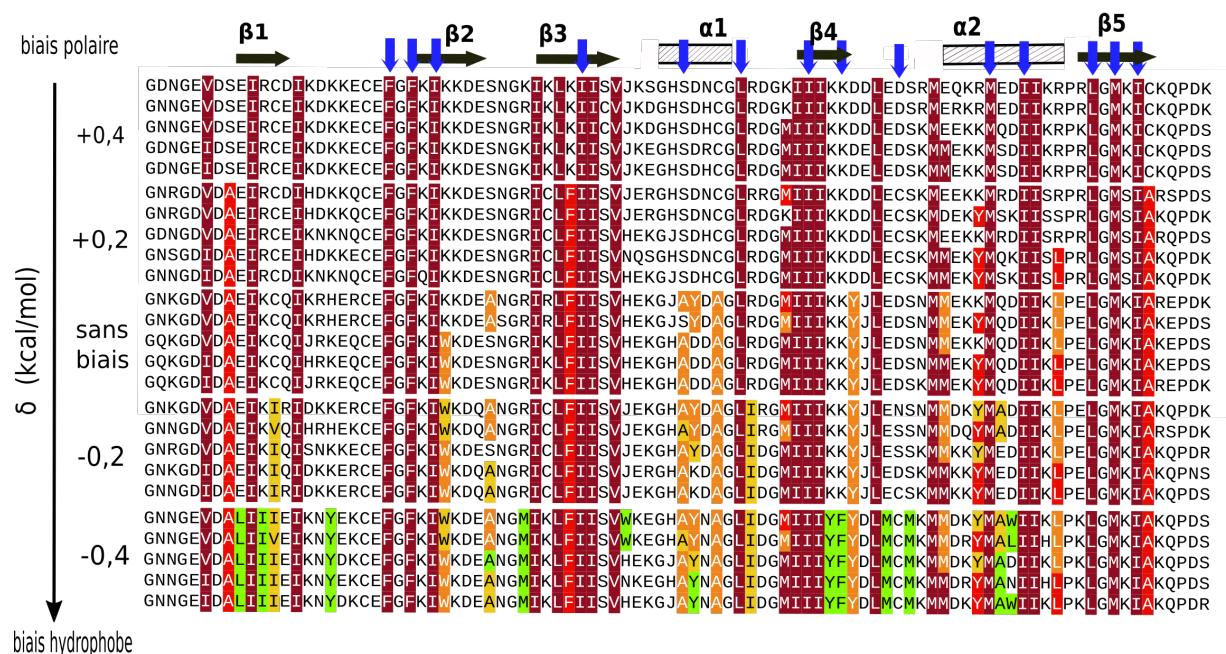


Figure 5.20 – Les séquences conçues par Proteus, les homologues à la native et les séquences naturelles représentées sous forme de logo pour les positions exposées

5.8 Application : Croissance du noyau hydrophobe

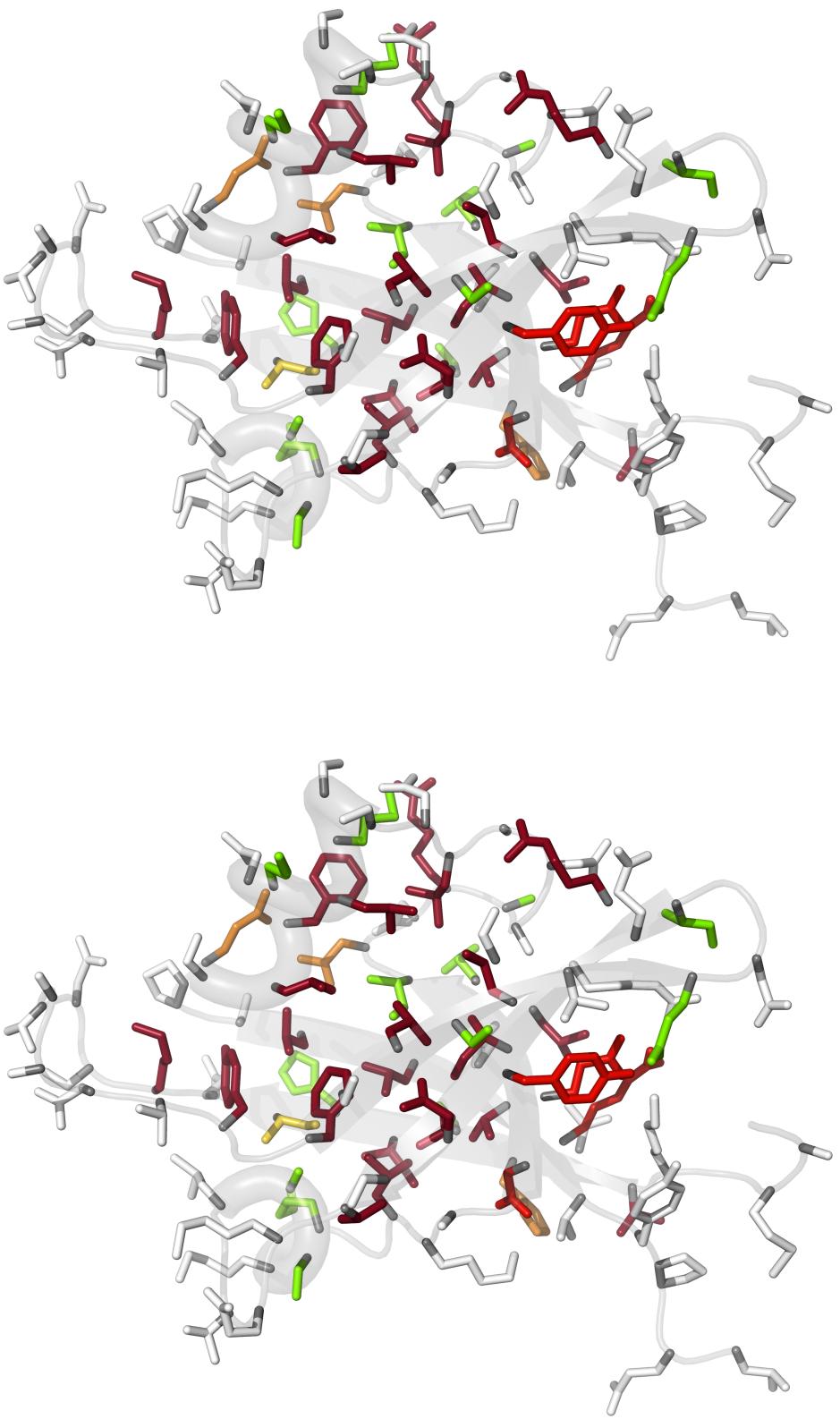
Comme application de nos modèles optimisés, nous examinons une possibilité de « design » du cœur hydrophobe des domaines PDZ. Deux de nos domaines PDZ, Tiam1 et Cask, sont soumis à une simulation REMC avec une succession de fonctions d'énergie biaisées qui favorisent progressivement les résidus hydrophobes. La première simulation comprend un terme d'énergie avec un biais $\delta = 0,4$ kcal/mol par position, qui pénalise les types d'acides aminés hydrophobes (I, L, M, V, A, W, F et Y). Le biais augmente alors graduellement et passe par les valeurs intermédiaires $\delta = 0,2$, $\delta = 0$ et $\delta = -0,2$ kcal/mol. La dernière simulation comprend un terme d'énergie de biais $\delta = -0,4$ kcal/mol (par position) qui favorise les types hydrophobes. En diminuant progressivement la valeur du biais d'énergie δ , nous « titrons » ainsi les résidus hydrophobes.

Figure 5.21 – Séquences Tiam1 obtenues avec un delta des énergies de références à -0,4, -0,2, 0, 0,2 et 0,4 et la structure native. Les hydrophobes pour des deltas de -0,4, -0,2, 0, 0,2 et 0,4 sont représentés par un dégradé allant du rouge foncé au vert clair en passant par le jaune.



Les flèches bleues indiquent les positions du cœur hydrophobe PDZ défini à partir de notre sélection de six domaines. Chaque groupe de séquences est une sélection à δ fixé parmi les séquences de plus faible énergie.

Figure 5.22 – Structure native Tiam1 avec les hydrophobes pour des δ de -0,4, -0,2, 0, 0,2 et 0,4 sont représentés par un jeu de couleurs allant du rouge foncé au vert clair en passant par le jaune.



Les résultats pour Tiam1 sont présentés à la figure 5.21 et à la figure 5.22. À la plus grande valeur de δ le cœur hydrophobe de Tiam1 est réduit, avec environ 10 positions d'acides aminés sur 94 qui changent en un type polaire, comparés aux séquences générées sans biais. Les positions modifiées se situent principalement sur le bord extérieur du cœur. À la valeur intermédiaire de 0,2 kcal/mol, le cœur hydrophobe ne compte plus que 4 ou 5 changements en type polaire. À la valeur de δ la plus négative, le cœur hydrophobe devient plus grand, s'étendant vers les régions de surface, avec globalement 14 positions polaires changées en types hydrophobes. Ainsi le nombre de positions modifiées est approximativement symétrique (environ +/- 12 changements), reflétant le biais. Environ 2/3 des changements se produisent dans des éléments de structure secondaire. Dans l'ensemble, les propensions observées de chaque position à devenir polaire ou hydrophobe en présence d'un biais de pénalité petit ou grand d'énergie δ peuvent être considérées comme un indice de design hydrophobe. Ici, 11 des 14 positions du cœur PDZ (toute sauf les positions 884, 898 et 903) sont restées hydrophobes au plus haut niveau de biais polaire, avec à peu près 13 autres positions, indiquant que ces positions ont la plus grande propension à être hydrophobes. De plus, près de 14 positions ont basculé de polaire à hydrophobe avec le biais le plus élevé, indiquant que ces positions aussi ont une certaine propension à être hydrophobes. Les résultats pour Cask sont similaires, avec 11 positions changées en polaire au plus haut biais polaire et 9 changées en hydrophobe au plus haut biais hydrophobe, voir 5.24.

Nous introduisons alors un indice pour décrire le nombre de changements relatifs de type d'acide aminé par unité d'énergie du biais. Cet indice ψ_h est défini comme le nombre δN de positions qui ont changées de non polaire à polaire, divisé par le produit de la variation δE dans l'énergie de polarisation et le nombre moyen N de positions non polaires à biais nul. Nous appelons ψ_h la sensibilité hydrophobe. Pour le domaine PDZ Tiam1, ce calcul donne : $\psi_h = \frac{1}{N} \frac{\delta N}{\delta E} = 0,9$ changements par position et par kcal/mol. Pour Cask, la sensibilité hydrophobe est $\psi_h = 0,7$ changements par position et par kcal/mol.

5.9 Conclusion

5.9.1 Modèle mis en œuvre

Nous avons paramétré notre modèle CPD pour la conception informatique de domaines PDZ, mis en œuvre dans le logiciel Proteus. Pour la modélisation de l'état replié, nous avons utilisé un champ de force protéique de qualité. Nous avons effectué des premiers paramétrages sur un ensemble de huit domaines PDZ, dont deux utilisés pour évaluer la

Figure 5.23 – Structure native Cask avec les hydrophobes pour des δ de -0,4,-0,2,0,0,2 et 0,4 sont représentés par un dégradé allant du rouge foncé au vert clair, en passant par le jaune.

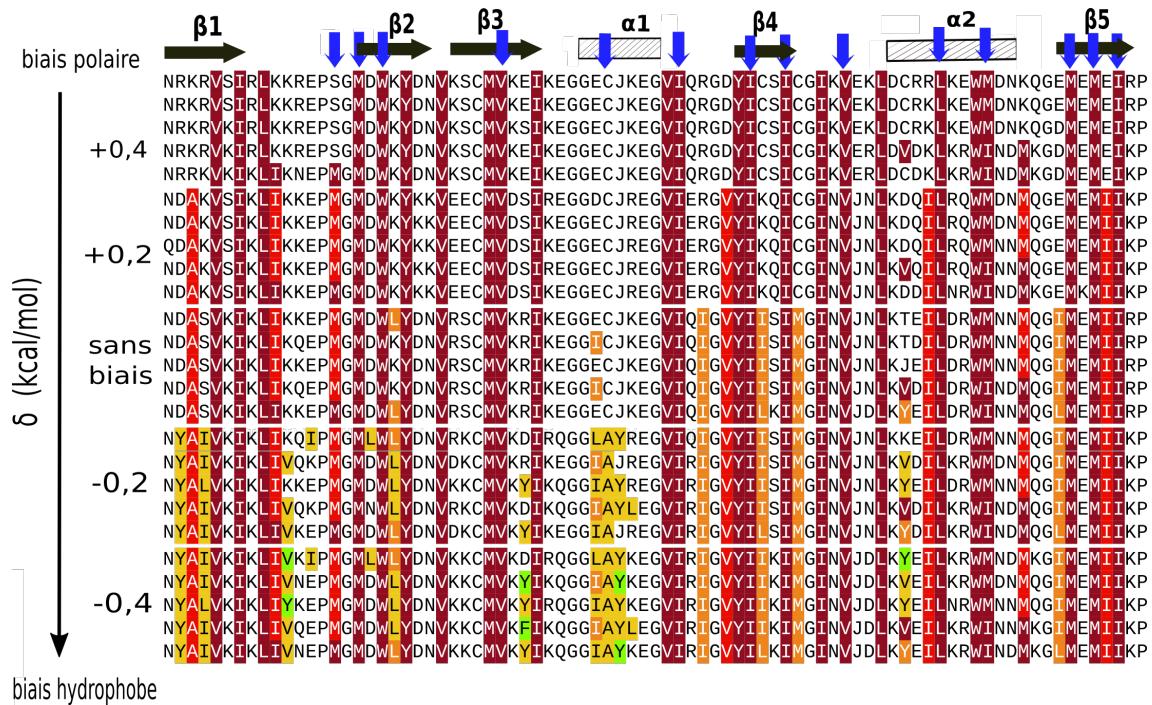


Figure 5.24 – Séquences Cask obtenues avec un delta des énergies de références à -0,4,-0,2,0,0,2 et 0,4 et la structure native. Les hydrophobes pour des deltas de -0,4,-0,2,0,0,2 et 0,4 sont représentés par un jeu de couleurs allant du rouge foncé au vert clair, en passant par le jaune.

transférabilité des résultats. Nous avons utilisé un premier traitement du solvant « GB », le modèle NEA et une constante diélectrique ϵ_P égale à 8. Puis nous avons effectué de nouveaux paramétrages sur un ensemble de trois domaines PDZ, avec une constante diélectrique ϵ_P égale à 4 et avec deux traitements du solvant : le NEA et la nouvelle méthode « GB », le modèle FDB et un jeu de paramètres surfaciques optimisé.

Pour les chaînes latérales, nous utilisons une bibliothèque de rotamères simple et discrète et une courte minimisation de chaque paire pendant le calcul de la matrice d'énergie, pour atténuer l'approximation de la discréétisation des rotamères. La fonction d'énergie et la description des rotamères ont été testées de manière approfondie et ont démontré de très bonnes performances pour les tests de reconstruction de chaînes latérales [59] (comparable au programme très populaire Scwrl4 [58]).

La représentation de l'état déplié utilise un modèle simple caractérisé par un ensemble de potentiels chimiques d'acide aminé empiriques ou énergies de référence. Ces énergies sont déterminées par une procédure de maximisation de vraisemblance, décrite ici, afin de reproduire la composition d'acides aminés d'homologues naturels soigneusement sélection-

nés. L'état déplié utilisé ici bénéficie d'un raffinement supplémentaire, puisque des valeurs d'énergies de référence distinctes sont utilisées pour les positions d'acides aminés selon qu'elles soient enfouies ou exposées à l'état replié.

Cette méthode suppose qu'il existe une structure résiduelle à l'état déplié, où certaines positions sont plus enfouies que d'autres. En outre, cela devrait rendre le paramétrage plus robuste et moins sensible à la taille et à la structure des homologues naturels utilisés pour définir les compositions d'acide aminé cibles, car les fréquences d'acide aminé des positions exposées et des régions enfouies sont calculées séparément. En principe, cela double le nombre d'énergies de référence à ajuster. Cependant, nous avons réduit ce nombre en introduisant des classes de similarités d'acide aminé, avec une seule énergie de référence ajustable par classe. Cette contrainte est levée dans la seconde moitié des cycles d'optimisation. Lors de l'optimisation des énergies de référence, nous effectuons, des calculs de séquences pour chaque protéine de notre jeu de test où une position sur deux peut muter (à l'exception de Gly et Pro), avec une simulation distincte pour chaque moitié. Ainsi, lors de l'optimisation des paramètres, une position mutable est toujours entourée d'un environnement identique au type sauvage au moins sur les deux positions immédiatement voisines sur le squelette. Les calculs s'appuient sur une méthode d'exploration Monte Carlo avec échange de réplique, qui utilise un demi-milliard de pas par simulation et produit des milliers de séquences par simulation.

Le modèle présente plusieurs limitations, dont la plupart sont très répandues en CPD. La première est l'utilisation de la stabilité des protéines comme seul critère de conception, sans prendre en compte explicitement la spécificité du pli [91, 32], la protection contre l'agrégation ou des considérations fonctionnelles comme la liaison des ligands. Toutefois, nous notons que les tests superfamily n'ont pas entraînés de mauvaises affectations (séquences perçues comme préférant un autre pli SCOP), donc en pratique, la spécificité du pli est vérifiée.

Une limitation supplémentaire est introduite par l'utilisation d'un squelette protéique fixe lors du calcul de la matrice d'énergie. En fait, le squelette n'est pas vraiment fixe. Certains mouvements sont autorisés, à travers l'utilisation d'une constante diélectrique protéique supérieure à 1 ($\epsilon_p = 4$ ou 8) [32]. Cette valeur diélectrique signifie que la structure protéique (y compris son squelette) est autorisée à se réorganiser en réponse à des mutations ou changements de rotamères. Cependant, la réorganisation est modélisée non pas explicitement, mais implicitement, et elle n'implique pas de mouvement des centres atomiques ou de leur sphère de Van der Waals associée. Ainsi, le squelette ne peut pas se réorganiser en réponse à une répulsion stérique produite par des mutations ou des changements de rotamères. L'utilisation d'un squelette fixe peut être en partie

compensée en concevant plusieurs structures PDZ. Par exemple, la mise en commun des séquences calculées sur six protéines a donné une entropie moyenne de séquence nettement plus proche de celle de l'ensemble expérimental Pfam. Une nouvelle méthode pour la conception de protéine multi-backbone a récemment été développée dans Proteus, sur la base d'une méthode Monte Carlo hybride qui préserve l'échantillonnage de la distribution de Boltzmann [117]. Cette méthode pourra être appliquée dans les prochaines études.

Une autre limitation de notre modèle est la nécessité, pour des résultats optimaux, de paramétriser les énergies de référence spécifiquement pour un ensemble donné de protéines. Cette étape est bien automatisée et de façon très parallèle. Cependant, cela implique plusieurs choix qui sont partiellement arbitraires. Ceux-ci comprennent le choix d'un ensemble de domaines protéiques pour représenter la protéine ou la famille d'intérêt. Nous devons également choisir un seuil de similarité pour définir les homologues cibles à partir desquels sont calculées les compositions expérimentales d'acides aminés. Ici, nous avons choisi d'utiliser les homologues de chaque membre de la famille, de calculer leurs compositions, puis de moyenner sur les familles. Cette méthode a bien fonctionné, mais d'autres choix sont possibles et des travaux complémentaires sont nécessaires pour pouvoir tirer des conclusions définitives sur ces choix.

5.9.2 Tests et application

Les séquences conçues par Proteus sont comparées aux séquences naturelles, à travers des tests de reconnaissance du pli, des calculs de similarité, des calculs d'entropie et des taux d'identité de séquences. Dans les simulations, nous concevons la totalité de la séquence de la protéine, de sorte que toutes les positions (à l'exception de Gly et Pro) peuvent muter librement, soumis à la seule contrainte de générer une composition moyenne en acides aminés similaire à la composition moyenne expérimentale (à travers les énergies de références). Malgré la quasi-absence de contraintes expérimentales, les séquences obtenues ont une forte similitude globale avec les séquences naturelles de Pfam, mesurée par les scores de similarité Blosum40. Les scores obtenus sont, pour l'essentiel, comparables aux scores de similarité entre les paires de séquences Pfam. La similitude est très forte pour les résidus au cœur de la protéine, comme cela a été observé dans des études CPD précédentes [90, 32]. En revanche, pour les résidus pris sur l'ensemble de la protéine, les scores de similarité sont plus faibles, mais l'approximation FDB couplée à une constante diélectrique ϵ_p égale à 4 donne toujours des séquences similaires à des homologues naturels modérément éloignés.

Notez que de nombreux résidus de surface sont impliqués dans des interactions fonctionnelles, comme les onze résidus de liaison aux peptides dans les domaines PDZ. Les résidus de surface sont également sélectionnés selon l'évolution pour éviter l'agrégation ou des adhésions indésirables. Ces contraintes fonctionnelles ne sont pas explicitement prises en compte dans notre protocole de design. Malgré ces difficultés sur les résidus de surface, la reconnaissance de pli avec l'outil Superfamily appliquée aux meilleurs modèles conçus est presque parfaite. Les tests de reconnaissance de pli antérieurs qui utilisaient une fonction d'énergie plus simple donnaient un taux de reconnaissance de pli inférieur, environ 85% (pour un ensemble de tests plus large et plus diversifié) et des similitudes inférieures [118, 81]. De toute évidence, l'utilisation combinée d'un champ de force protéique amélioré, du solvant GB FDB et des énergies de référence spécifiques à la famille conduisent à des séquences calculées proches des séquences natives.

Les séquences Proteus ont également été comparées aux séquences obtenues avec le logiciel Rosetta, qui a lui-même été testé de manière approfondie. Sur la base des scores de similarité de Blosum (par rapport aux séquences naturelles dans Pfam) et des tests de reconnaissance du pli, les séquences Proteus et Rosetta sont globalement de même qualité. Cependant, Rosetta fait moins de mutations que Proteus ; de sorte que les scores d'identité, par rapport à la protéine de type sauvage correspondante, sont entre 7% et 9% plus haut chez Rosetta que pour la version FDB de notre modèle. Ce qui veut dire que Proteus modifie environ cinq positions en plus, en moyenne, par domaine PDZ.