# Assignment of secondary structures of proteins : DSSP

Karim Bouchaara
M2 Bioinformatics, Université Paris Cité

2025-2026

# Introduction

Assigning protein secondary structure (SS) from 3D coordinates is a staple of structural bioinformatics: it compresses folds into a 1-D code, enables comparisons across homologs, supports benchmark construction for prediction, and underlies residue-level analyses. The community standard for this task is DSSP ("Dictionary/Define Secondary Structure of Proteins"), introduced by Kabsch and Sander (Biopolymers, 1983). DSSP labels each residue using physically grounded rules: a peptide–peptide hydrogen-bond (H-bond) energy combined with simple geometric/motif criteria yields helices (3-10, α, π: G/H/I), β-structure (B/E), turns (T), bends (S), and coil. Because many downstream evaluations collapse to the coarse H/E/C alphabet, an assignment that is transparent and reproducible remains highly valuable.

At the heart of DSSP is an electrostatic model for backbone H-bonds between carbonyl CO and amide NH groups. Partial charges on C=O and N–H define an energy $E = 332 \cdot q_1 \cdot q_2 \cdot (1/r\_ON + 1/r\_CH - 1/r\_OH - 1/r\_CN)$ with $q_1 = 0.42$, $q_2 = 0.20$ (distances in Å; energy in $\text{kcal} \cdot \text{mol}^{-1}$); more negative E indicates a stronger bond. From these bonds, DSSP detects repeating turns and bridges that cooperatively build regular elements. A turn of order n is CO(i)→NH(i+n); two consecutive n-turns seed a helix of type G (n=3), H (n=4), or I (n=5), which is then extended by overlap. β-structure begins with local β-bridges between residues on two strands; specific pairings distinguish parallel from antiparallel bridges, and consecutive bridges of the same type form a β-ladder. Residues in ladders are annotated as strands (E), whereas isolated bridges receive B. A fixed precedence (H > B > E > G > I > T > S) resolves overlaps so that each residue gets a single symbol. This principled, deterministic scheme became a standard because it is objective, reproducible, and maps naturally to H/E/C (folding G and I into H, B into E).

In this work we re-implement the core of the 1983 DSSP in a compact Python program and adhere to those original definitions. Starting from PDB coordinates, we use backbone N, Cα, C, O, add a virtual amide H when missing, compute DSSP-style CO···NH energies and declare bonds when E passes a negative threshold (default $-0.45\ \text{kcal} \cdot \text{mol}^{-1}$), detect n-turns and derive helices from consecutive turns, identify parallel and antiparallel β-bridges and merge them into ladders to assign strands versus isolated bridges, then apply the DSSP priority to produce one SS string per chain. We evaluate against the official mkdssp on three small proteins (1CRN, 1BTA, 1UBQ), report three-state accuracy (Q3 over H/E/C, with G/I→H and B→E), probe sensitivity to the energy cutoff, and provide a per-residue qualitative comparison on a representative example. Throughout, we critically examine our choices—especially virtual-hydrogen placement, a single global cutoff, and simplified β rules—and discuss divergences from mkdssp and straightforward improvements (amide-plane H placement, stricter helix cores, inclusion of bends and solvent accessibility).

# Materials and Methods

We implemented a compact DSSP-like assigner in Python (NumPy, Biopython) that follows the original Kabsch–Sander rules for regular secondary structures while keeping the code minimal and transparent. Input structures are read from PDB files with Biopython's PDBParser (quiet mode). We analyze a single model (structure[0]) and a user-specified chain (default "A"). Only standard amino-acid residues are retained (res.id[0] == " "); HETATM and waters are skipped. For each kept residue we require the backbone atoms N, Cα, C, and O to be present; when available we also record H (or HN) on the backbone nitrogen. Residues that miss any of N, Cα, C, or O are dropped from the analysis.

Because many PDBs lack amide hydrogens, our pipeline places a virtual H on N when needed. If the previous peptide carbonyl C is plausibly connected to the current N (C–N distance within 1.2–2.2 Å), we construct H using the bisector of the vectors N→(prev C) and N→Cα; otherwise, we fall back to the bisector of N→C and N→Cα (first residue or chain break). In both cases we normalize the direction and set |N–H| ≈ 1.0 Å. This local-geometry placement is simple and robust enough to define all interatomic distances required by the DSSP energy. For numerical safety we clamp any distance used in energy calculations to a minimum of $1\times10^{-6}$ Å.

Hydrogen bonds (H-bonds) are scored with the classic DSSP electrostatic model on peptide groups: partial charges q1 =0.42 (on the C=O pair) and q2 =0.20 (on the N–H pair) enter

$E=332q1\ q2\ (rON\ 1 +rCH\ 1 -rOH\ 1 -rCN\ 1\ ),$

with distances rXY in ångströms and energy in kcal·mol$^{-1}$. For every residue pair (i,j) we evaluate the energy of CO(i) against NH(j), skipping near neighbors with $|i-j|<2$ to avoid trivial local contacts. A pair is considered an H-bond if E is strictly below a configurable cutoff (default −0.45 kcal·mol$^{-1}$; sensitivity explored at −0.40 and −0.55). The set of accepted pairs is stored as a dictionary (i,j)↦E where (i,j) denotes CO(i)→NH(j).

Helical states are derived from n-turns, defined as bonds CO(i)→NH(i+n) for $n \in\{3,4,5\}$. We first collect all indicesi that realize a 3-turn, 4-turn, or 5-turn. For each n, two consecutive n-turns at i−1 and i seed a minimal helix whose span corresponds to the helix type: 3–10 (G, n=3, minimal span i..i+2), α (H, n=4, span i..i+3), or π (I, n=5, span i..i+4). Helices are then extended by overlap: as long as the window of consecutive n-turns advances by one position, the helix end is lengthened accordingly. This rule suppresses single, noisy turns while preserving the cooperative nature of helices.

β-structure is built from β-bridges and β-ladders. We scan residue pairs (i,j) (excluding $|i-j|<3$) and recognize a parallel bridge (P) if we find the pattern (i−1,j) and (j,i+1) among accepted H-bonds (symmetrically, (j−1,i) and (i,j+1) also qualify). We recognize an antiparallel bridge (A) if
(i,j) and (j,i) are both present, or if (i−1,j+1) and (j−1,i+1) are both present. Bridges of the same type are then chained into ladders by stepping forward along the strands: (+1,+1)
(+1,+1) for parallel and (+1,−1) for antiparallel, with symmetric backward extension, until the pattern breaks. Residues that participate in ladders of length at least two bridges are labeled as extended strands (E); residues involved only in a single, isolated bridge are labeled B.

Non-helical turns (T) are assigned at the endpoints of detected n-turns: for each turn at i→i+n, we mark positions i and i+n as T only if they are still unassigned (coil) after helix and β assignment. Finally, to resolve overlaps we apply a DSSP-like priority on a per-residue basis: H > B > E > G > I > T > S > C. This yields a single-letter secondary-structure string for the chain. The program also prints, on request, residue names and counts, the strongest hydrogen bonds, and the list of detected bridges for qualitative inspection.

The command-line interface accepts the PDB path, a chain identifier (default A), the H-bond energy cutoff (default −0.45 kcal·mol$^{-1}$), and flags for verbose output, bridge lists, and top H-bonds. Runtime scales roughly as O(L^2) in chain length due to the all-pairs CO···NH scan and the bridge search, which is negligible for the small test proteins used here. All evaluations in the Results section compare our assignment to a reference DSSP produced by mkdssp on the same PDB files; for three-state accuracy (Q3) we map G/I→H and B→E on both sides, and aggregate everything else into C.

# Results

We evaluated our mini-DSSP against the mkdssp reference on three proteins (1CRN, 1BTA, 1UBQ). For each chain we report the composition of our prediction in H/E/C and the three-state accuracy Q3 after mapping G and I to H and B to E. The agreement is consistently high: 1CRN reaches Q3 = 91.3% with predicted composition 43.5% H, 8.7% E, 47.8% C; 1BTA attains Q3 = 96.6% with 41.6% H, 18.0% E, 40.4% C; and 1UBQ scores Q3 = 100.0% with 15.8% H, 31.6% E, 52.6% C. The mean Q3 over all three proteins is 96.0%, indicating that the core DSSP logic (H-bond energy + motif rules) is well captured by our implementation (Figure 1).

| PDB | L | %H | %E | %C | Q3 |
|---|---|---|---|---|---|
| 1CRN | 46 | 43.5 | 8.7 | 47.8 | 91.3 |
| 1BTA | 89 | 41.6 | 18.0 | 40.4 | 96.6 |
| 1UBQ | 76 | 15.8 | 31.6 | 52.6 | 100.0 |
| **Mean** | — | — | — | — | **96.0** |

**Figure 1 :** Per-protein H/E/C composition and Q3 of mini-DSSP vs mkdssp

We compared our assignments to the DSSP reference in H/E/C after mapping G and I to H and B to E, and we summarize the agreement with confusion matrices for each protein. In 1CRN (Figure 2; rows = reference DSSP, columns = prediction), the overall agreement is high (Q3 = 91.3%), with β-strands recovered perfectly (4 true positives, no errors) and most discrepancies occurring at helix boundaries as coil↔helix flips (H: 19 TP and 3 FN; C: 19 TP and 1 FP). In 1BTA (Figure 3), the match is even stronger (Q3 = 96.6%): helices and strands are exact (37 TP H, 16 TP E), and the only three errors are coil residues predicted as helix at helix termini, consistent with borderline hydrogen bonds near the edges. In 1UBQ (Figure 4), the agreement is complete (Q3 = 100.0%), with no off-diagonal counts after the H/E/C mapping. Taken together, these three figures show a consistent pattern: β-sheets are robustly captured, while the few residual disagreements arise where expected—at the starts and ends of helices—echoing our cutoff-sensitivity analysis.

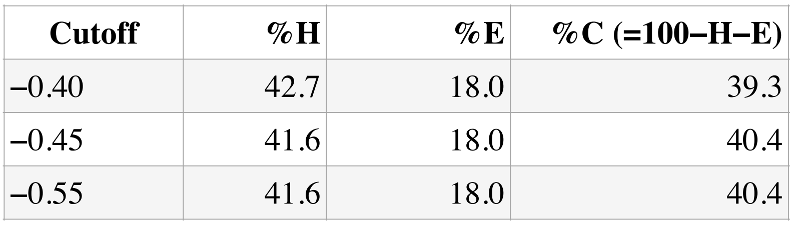
| **Ref \ Pred** | **H** | **E** | **C** |
|---|---|---|---|
| **H** | 19 | 0 | 3 |
| **E** | 0 | 4 | 0 |
| **C** | 1 | 0 | 19 |

| **Ref \ Pred** | **H** | **E** | **C** |
|---|---|---|---|
| **H** | 37 | 0 | 0 |
| **E** | 0 | 16 | 0 |
| **C** | 3 | 0 | 33 |

| **Ref \ Pred** | **H** | **E** | **C** |
|---|---|---|---|
| **H** | 18 | 0 | 0 |
| **E** | 0 | 26 | 0 |
| **C** | 0 | 0 | 32 |

**Figures 2, 3 & 4 :** H/E/C confusion matrices (rows: DSSP reference; columns: mini-DSSP) for 1CRN, 1BTA, and 1UBQ.

Varying the DSSP-style cutoff from −0.40 to −0.55 kcal·mol$^{-1}$ leaves the β-strand content essentially unchanged at 18.0%, indicating that sheet ladders are robust to modest threshold shifts. The helical fraction changes only slightly (42.7% → 41.6%), with a compensating change in coil (39.3% → 40.4%), consistent with marginal bonds at helix termini turning on/off around the threshold. We therefore keep −0.45 kcal·mol$^{-1}$ as the default for the remainder of the analysis (Figure 5).



| Cutoff | %H | %E | %C (=100–H–E) |
|---|---:|---:|---:|
| –0.40 | 42.7 | 18.0 | 39.3 |
| –0.45 | 41.6 | 18.0 | 40.4 |
| –0.55 | 41.6 | 18.0 | 40.4 |

**Figure 5 :** Sensitivity of H/E/C composition to the hydrogen-bond energy cutoff on 1BTA

# Discussion

Our mini-DSSP reproduces the core logic of DSSP—electrostatic CO $\cdots$NH scoring plus explicit motif rules for n-turns and β-bridges—and achieves strong three-state agreement (mean Q3 = 96%) on three small proteins. The β-strand content is remarkably stable across reasonable H-bond energy thresholds, reflecting that sheet ladders are supported by multiple, mutually reinforcing bridges. Residual disagreements concentrate at helix boundaries, where borderline hydrogen bonds can appear or disappear with small geometric fluctuations; this explains the small H⇔C flips seen in the confusion matrices and in the cutoff sensitivity. The high accuracy on 1UBQ (Q3 = 100%) indicates that, for well-resolved structures, a simple virtual-hydrogen model and the classic DSSP energy suffice to recover H/E/C without ad-hoc tuning.

The approach has limitations inherent to its minimalism. Virtual amide hydrogens are placed from local bisectors rather than an explicit amide plane, which can perturb rOH and rCH in crowded regions. A single global energy cutoff cannot capture context (e.g., resolution, B-factors, proline effects), and our β-bridge rules consider only the bond pattern, not detailed angular criteria. We also omit DSSP's bends (S) and solvent accessibility, which can refine assignments in loops. Finally, ladders are built as disjoint runs, so shared strands or forked topologies are simplified.

These constraints suggest straightforward extensions. A more faithful H placement (amide-plane construction with idealized peptide geometry) and a slightly stricter helix core (e.g., require three consecutive n-turns before labeling H) should reduce boundary errors. Incorporating bends and basic solvent accessibility would improve interpretability in coils. A distance prefilter (KD-tree) would lower the

O(L^2) cost for larger chains without changing results. Even with these omissions, the present implementation is reproducible, deterministic, and pedagogically useful: it exposes exactly how DSSP's patterns give rise to helices and sheets and why β content is robust while helix termini are sensitive.

# Conclusion

We implemented a compact, DSSP-style assigner that captures the core of Kabsch–Sander electrostatic CO $\cdots$NH scoring plus explicit n-turn and β-bridge rules—and validated it on three proteins with a mean three-state accuracy Q3 = 96% against DSSP. β-strands are consistently recovered and remain stable under reasonable changes of the hydrogen-bond cutoff, while the small discrepancies concentrate at helix boundaries, in line with the expected sensitivity of edge H-bonds. Despite simplified hydrogen placement and β criteria, the method reproduces DSSP closely and provides a transparent, reproducible baseline. Future improvements will target amide-plane H placement, stricter helix cores, inclusion of bends and solvent accessibility, and richer evaluation (e.g., per-class metrics and segment overlap) on larger benchmarks.