Problem 5.1 [Consecutive Sets]. Prove that the following problem is NP-complete.

CONSECUTIVE SETS: Given a collection of (unordered) subsets $S_1, S_2, \ldots, S_n$ of a finite alphabet $\Sigma$, and a positive integer $k$, is there a string $w$ over the alphabet $\Sigma$ with length at most $k$ such that, for each $S_i$, the elements of $S_i$ occur (in any order) as some consecutive characters $w_j, w_{j+1}, \ldots, w_{j+|S_i|-1}$ of $w$?

Hint: Reduce from some version of Hamiltonicity.

Solution: We first prove that CONSECUTIVE SETS is in NP by giving a nondeterministic polynomial-time algorithm to decide it. If $k \geq \sum |S_i|$ then we immediately return YES, since the trivial solution which simply lists all the subsets is a solution. Otherwise, we nondeterministically guess a string $s$ of length $k$. Then for each $i$ we nondeterministically guess an offset and a permutation of the elements of $S_i$, and verify that those elements indeed appear in $s$ at that offset in order. If this verification succeeds for all of the subsets, then we return YES; otherwise we return NO.

This algorithm requires linear time to guess the string $s$, and linear time to verify that each subset appears in $s$. Thus it takes at most quadratic time to check all of the subsets, so it is polynomial-time as desired. Therefore CONSECUTIVE SETS is in NP.

We now prove that CONSECUTIVE SETS is NP-hard by reducing from HAMILTONIAN PATH in SIMPLE 3-REGULAR UNDIRECTED GRAPHS. Let $G = (V, E)$ be a simple, 3-regular, undirected graph. Let $\Sigma = E$ and for each vertex $v \in V$ let $S_v$ be the set of edges adjacent to $v$. Finally, set $k = 2|E| - (|V| - 1)$. We output the CONSECUTIVE SETS instance $(\Sigma, S_v, k)$. This reduction is $O(|E|)$ (it includes each edge twice); thus it is polynomial-time.

Note that

$$S_u \cap S_v = \begin{cases} (u, v) & \text{u is adjacent to v} \\ \emptyset & \text{u is not adjacent to v} \end{cases}$$

for distinct vertices $u, v$.

We now show that our reduction is correct. Suppose that there exists a Hamiltonian Path on $G$, which visits the vertices in order $v_1, \ldots, v_n$. Then there exists a solution to the corresponding CONSECUTIVE SETS instance which is obtained by concatenating the sets $S_{v_1}, \ldots, S_{v_n}$, overlapping each adjacent pair $S_u, S_v$ using the edge $(u, v)$. The resulting string has length $2|E| - (|V| - 1) = 2|V| + 1 = k$, since each edge is output twice except that we overlap $|V| - 1$ pairs of them. Therefore, it is a solution to the CONSECUTIVE SETS instance.

Conversely, suppose that there exists a solution to the CONSECUTIVE SETS instance; that is, a string $w$ of length at most $k$ which contains each of the $S_v$. Because each pair of $S_u, S_v$ have intersection of size at most 1 and each $|S_v| = 3$, each subset can overlap with at most two others, and by a margin of only 1 character. Thus such a $w$ must have length at least $3|V| - (|V| - 1) = 2|V| + 1 = k$, since it includes $|V|$ subsets of size 3 and we can save only $|V| - 1$ characters by

\footnote{Formally, we create an alphabet with a symbol for each edge, but we omit this layer of indirection for clarity.}
overlapping. Thus \( w \) has length exactly \( k \). Therefore every subset overlaps by 1 character with the subsets next to it in the string. But this implies that the corresponding vertices are adjacent. Thus there exists a chain of \(|V|\) vertices including every vertex, where every vertex is adjacent to those next to it in the chain. This is exactly a Hamiltonian Path in \( G \).

Therefore our reduction is sound, showing that Consecutive Sets is NP-hard. Because Consecutive Sets is NP-hard and in NP, it is NP-complete.